UNIVERSITY OF CALIFORNIA
SANTA CRUZ

IDENTIFYING KEY PATHWAYS IN MULTIPLE CANCERS WITH MULTI-OMICS PATHWAY ANALYSIS

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

DOCTOR OF PHILOSOPHY

in

BIOMOLECULAR ENGINEERING & BIOINFORMATICS

by

Sam Ng

June 2015

The Dissertation of Sam Ng is approved:

____
Professor Josh Stuart, Chair

____
Professor David Haussler

____
Professor Christopher Benz

____
Professor Eric Collisson

____
Tyrus Miller, Ph.D.
Vice Provost and Dean of Graduate Studies

Table of Contents

List of Figures

Abstract

IDENTIFYING KEY PATHWAYS IN MULTIPLE CANCERS WITH MULTI-OMICS PATHWAY ANALYSIS

By

Sam Ng

Since response to therapy can differ greatly between cancer patients, a precision medicine approach to treating cancer based on the uniqueness of patient tumors could greatly improve response rate and quality of life. High-throughput assays provide the means to probe multiple types of genomic alterations across a patient's cancer genome. By leveraging prior knowledge about genetic pathways, I have created tools to address the challenges of tackling large multi-dimensional datasets to make biological sense of the data. I developed *PATHMARK* to identify clusters of genes that are dysregulated together forming networks that offer insights into disease mechanisms and treatment strategies. *PATHMARK* utilizes conventional univariate differential analysis with a filter on pathway interactions to identify sub-networks that are significantly more connected than by chance. I developed *PARADIGM-SHIFT* to predict the functional impact of mutations detected from whole-exome

sequencing data. *PARADIGM-SHIFT* analyzes the inferred activities of a network surrounding a mutated gene, comparing the levels between mutant and wild-type samples. The approach predicts if mutations are likely to be neutral, gain-of-function, or loss-of-function. I demonstrate how inferences about mutations in novel genes and in non-coding regions can be gleaned from models trained on known coding mutations. The predictions form “molecular machines” that link events together based on shared pathway alteration. With the growing number of available datasets and computational tools, it has become increasingly important to make analyses easily accessible and reproducible. To support these ideals, I have developed my tools to be compatible within the Galaxy system, which has enabled collaborators to apply my tools to analyze their data.

Acknowledgements

I would like to thank my committee, my numerous supportive instructors, my colleagues here in the UCSC Cancer Genomics Group, my many collaborators at Five-3 Genomics, The Buck Institute, UCSF, The Cancer Genome Atlas project, my friends, my family, and my wife Ann Chang. I could not have made it thus far without your academic and moral support.

1 Introduction and Overview

The management – diagnosis, prognosis, and treatment – of cancer has evolved tremendously in the past few years. Our understanding of the molecular differences and similarities between tumors across thousands of patients originating from many tissue types has grown due to the advent of lower cost genomic assays. These assays allow researchers and clinicians to probe many facets of a patient's tumor and make observations not visible by examining tumor cells under a microscope. Within the tens of thousands of data points for each patient's tumor is a wealth of knowledge, extracting that knowledge remains a major challenge. This is especially true due to the "curse of dimensionality," the problem that arises when the number measurements per sample is much larger than the size of the cohorts being studied. Because of this high dimensionality, it is difficult to determine which features are truly informative and which ones appear informative by chance. Large cohort studies such as the TCGA (The Cancer Genome Atlas) project have contributed greatly to our growing knowledge base, though this alone is not enough. Recent efforts to incorporate prior biological knowledge through the incorporation of pathway knowledge have been very successful for biological discovery and making connection to biological mechanisms. Many of these improvements have been achieved through the use of pathway gene sets: groups of related genes known to be involved together to perform some function, though some methods also utilize the knowledge of which genes directly interact and how they affect each other directly within a cell. In this thesis, I will outline the tools that I have developed for analyzing cancer genomics data with the goal of improving treatment decisions and biological discovery.

In Chapter 1, I discuss the effects genomics has had on our understanding and treatment of cancer. Computational methods can utilize molecular profiles derived from patient tumors to guide treatment decisions on predicted drug response. I show that integrating pathway knowledge into these computational methods improves predictions and introduce *PARADIGM* an algorithm for integrating genomics data into a graphical pathway model used as a foundation for much of my own method development.

In Chapter 2, I demonstrate how pathway topology can be integrated with traditional biomarker discovery in our *PATHMARK* analysis. *PATHMARK* has successfully identified pathways that predicted drug response within cancer subtypes, as well as improved biological interpretability.

In Chapter 3, I present *PARADIGM-SHIFT* a method for predicting the functional impact of mutations using genomic data and pathways. With the increased prevalence of tumor-specific mutation annotations due to the access to high-throughput whole-exome sequencing, a growing number of recurrent mutations were identified in many different cancer types. However, understanding the impact of missense mutations is difficult. Building from *PATHMARK*, *PARADIGM-SHIFT* was developed to determine whether these driver mutations lead to a gain-of-function (GOF) or loss-of-function (LOF) by examining the effect of mutation on pathway signaling.

In Chapter 4, I present several surprising findings stemming from applying *PARADIGM-SHIFT* to several TCGA cohorts. *PARADIGM-SHIFT* was not only able to predict the functional impact of mutation for many highly recurrent alterations, but useful for identifying events that lead to the gain- or loss-of-function of a particular pathway. At first, known events were incorporated to look for support of a similar *PARADIGM-SHIFT* prediction across events, then I developed a method for discovering these events by looking for associations between the predicted functional impacts and alterations in genes likely to be involved in the same pathway.

In Chapter 5, I address the need for computational tools to be accessible and reproducible, highlighting my work in making my tools available through the web and as a module within *Galaxy*. *Galaxy* is a publically available, web-based platform for running computational tools that enables researchers to address these specific needs.

Through the course of my research, I have contributed results to many working groups and as co-authored several papers, in addition to my accepted methods paper to the ECCB '12 edition of *Bioinformatics*. This list includes the many papers I have been involved with:

- Subtype and pathway specific responses to anticancer compounds in breast cancer. *PNAS* February 2012
- Whole-genome analysis informs breast cancer response to aromatase inhibition, *Nature* June 2012
- Comprehensive molecular characterization of human colon and rectal cancer, *Nature*, July 2012
- Comprehensive genomic characterization of squamous cell lung cancers, *Nature*, September 2012
- PARADIGM-SHIFT predicts the function of mutations in multiple cancers using pathway impact analysis, *Bioinformatics*, September 2012

- Comprehensive molecular portraits of human breast tumours, *Nature*, October 2012
- Integrated genomic characterization of endometrial carcinoma, *Nature*, May 2013
- Comprehensive molecular characterization of clear cell renal cell carcinoma, *Nature*, July 2013
- The UCSC Interaction Browser: multidimensional data views in pathway context, *Nuclear Acids Research*, July 2013
- Comprehensive molecular profiling of lung adenocarcinoma, *Nature*, July 2014
- Multiplatform analysis of 12 cancer types reveals molecular classification within and across tissues of origin, *Cell*, August 2014
- Comprehensive molecular characterization of gastric adenocarcinoma, *Nature*, September 2014
- TCGA melanoma submitted
- TCGA lower grade glioma submitted

I have contributed figures to many of the papers listed here and the results will be covered in further detail in the following chapters.

1.1 Precision Medicine in Cancer

Cancer is not a single disease. The variations from one tumor to another can result in vastly different responses to treatment. Due to this heterogeneity, strategies for determining the best treatment options available to a particular patient are necessary for advancing cancer treatment. With the rise of –omics data availability, researchers are able to probe many of the molecular facets of a particular cancer and use that information to guide treatment and make treatment decisions based on their unique molecular characteristics. As a result, newer treatments have been developed that target specific weaknesses of certain tumors in addition to traditional treatments, such as surgery, radiation therapy, and chemotherapy. This provides an alternative perspective for clinicians to maximize treatment effectiveness, while also reducing the burden of unsuccessful treatments and side effects.

One such therapy is the successful anti-*HER2* (*ERBB2*) monoclonal antibody trastuzumab (Herceptin). Developed by Genentech, trastuzumab has shown great success in treating *HER2* overexpressing tumors, one of the major subtypes in breast cancer. While, trastuzumab shows tremendous clinical efficacy in this subset of breast cancer patients, other patients would not benefit from this treatment, which instead would be a burden on those individuals. Therefore, there are tests to determine whether or not a patient is likely to benefit from treatment by assessing the overexpression or amplification of *HER2*. This is an example of a biomarker-guided therapy, in which the presence of a particular marker indicates likelihood of response.

While trastuzumab has shown great success with certain tumors, there are many patients with tumors that do not have good biomarkers to inform treatment. For example, in lung adenocarcinoma, roughly a third of the patients are considered to be oncogene-negative; i.e. lack a known cancer driving GOF mutation. This could mean that there are no effective treatments developed that the patient will respond to or that no biomarker has been discovered to indicate the use of some drug that would be effective. In either case, there is much unknown about tumor diversity and how to optimally treat these tumors. In recent years, several large cohort studies headed by the NCI (National Cancer Institute) involving research groups all around North America have been at work under the TCGA (The Cancer Genome Atlas) project. The goal of the TCGA project has been to bring together one of the largest collections of tumors across many different tumor types and to characterize them with a full suite of genomic assays. Pathway-based analyses have been very successful in making novel discoveries in the TCGA projects and many methods, including the ones I will outline in this thesis, have been developed and applied to these datasets.

1.2 Driver and Passenger Events in Cancer Evolution

As an individual ages, their somatic cells accumulate mutations, these could be due to exposure to mutagens or errors in biological machinery. The majority of these mutations are harmless, *passenger* mutations that merely accumulate and have little impact on the fitness of the cells, but occasionally a *driver* mutation may occur leading to a selection advantage for the developing tumor. As this population expands it acquires new *passenger* and new *driver* mutations that eventually lead to cancer. With such a wide array of different kinds of mutations possible, from large copy number amplifications, deletions, and gene fusions to small point mutations making the distinction between *passenger* and *driver* is a difficult challenge and a distinction that may not be completely black and white, but very much depend on the unique context of each tumor. The presence of multiple *driver* mutations may require a combination of therapies in order to suppress the tumor, and the unique variability within a tumor through the accumulation of *passenger* mutations may lead to the development of treatment resistance as the tumor adapts to escape cell death.

Researchers and clinicians have only scratched the surface on understanding the mechanism(s) underlying *driver* mutations in cancer. Many rare mutations are not well understood and may require unique treatments to combat them. Only by observing more cases and incorporating additional knowledge can we hope to understand and one day treat these tumors driven by rarely, but important genes.

1.3 How Pathway Analysis Can Improve Cancer Treatment

In real biological settings, genes do not work in isolation, but interact with each other to achieve some function. As such, we should not treat them as separate features when developing our computational methods. By leveraging these genetic interactions, our models are improved by taking into account information that is external to our data and builds upon a foundation of knowledge from past research studies. The type of pathway information used can vary from gene modules, or sets of related genes known to be involved together, to direct genetic interactions that describe how genes and their protein products interact with each other in a cell.

Pathway knowledge in the form of gene-to-gene interactions can empower us to make discoveries by bringing in orthogonal information not present in our datasets. A result from the TCGA glioblastoma multiforme (GBM) characterization paper identified that while individual alterations in genes do not appear significantly recurrent, such as RAS mutations occurring in only 2% of patients, when taken together the total percentage of alterations across these pathways is significantly recurrent and occur in a majority of patients. By grouping these genes together it makes the result more interpretable as well, many of the genes altered in these pathways are involved with cellular processes that would give the tumors a selective advantage. For example, the RTK/RAS/PI(3)K signaling pathway is altered in 88% of patients allowing tumors to escape apoptosis. Without pathway knowledge many of these alterations would go undetected, pathways are a useful tool for improving interpretability of results and increasing statistical power of our analyses.

1.3.1 Pathway Biomarkers of Therapeutic Response

Overexpression is not the only mechanism for increasing the total activity of a gene. In the case where a treatment is effective at shutting down the function of an oncogene that a tumor requires to survive, patients who may respond to therapy could be missed if only looking for overexpression. A mutation could lead to over activity of the protein rather than appearing as protein overexpression, and this gain of function could be detected by looking at the effect of the mutation on the pathway. Alternatively, mutations in genes within the pathway may alter the pathway and might be missed by single gene approaches. Biomarkers could be expanded to look for pathway effects rather than single gene activities.

1.3.2 The Challenge of Apparent Oncogene Negative Tumors

Many of the frequent *driver* mutations have already been characterized and some of the oncogenes have successful targeted therapies to treat them, however, there are still many patients that do not have mutations in any of these well-characterized oncogenes. This is a major challenge because at lower mutation frequencies there is much less statistical power for determining the importance or impact of gene mutations as well as a large quantity of these uncharacterized mutated genes. Discovering these genes is essential for designing therapies that will be effective for these patients.

1.4 PARADIGM an Integrative Approach for Understanding Cancer

What matters, with a few expectations of lncRNAs is protein activity. When we talk about mutated genes, we are really referring to the sequelae of those genes' protein products. Furthermore, protein activities are only rarely measured directly, but instead are inferred from surrounding measurable consequences. Take for example a metabolic enzyme that acts on a substrate to produce a product. One could argue that the only definitive way to assess the activity of the enzyme is to measure the rate of substrate turnover. However, if a direct measure of the disappearance of the substrate or appearance of the product is not available then one could look at the activity of other enzymes that depend on the product as their substrate to infer the activity. This leads to a recursive definition of the activity of the enzyme in terms of the activity of other neighboring enzymes. In the same way, we can infer the activity of a transcription factor or kinase based on the activities of other genes in their regulatory neighborhoods.

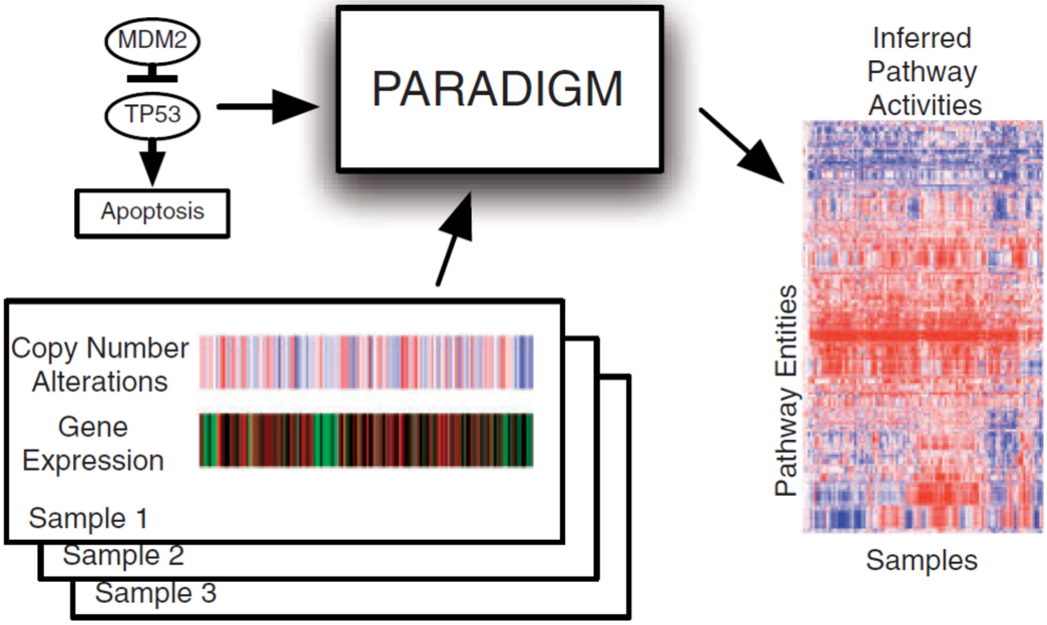


Figure 1 *PARADIGM* model of integrating genomic data with a pathway model to infer pathway activities.

PARADIGM is a factor-graph-based approach for integrating diverse types of omics data with genetic pathways, Figure 1 [Vaske 2010]. The approach assesses the activity of a gene in the context of a genetic pathway diagram ϕ by drawing inferences from a dataset of observations D. The dataset can include multiple different types of measurements for a patient sample such as gene expression and genomic copy number variation. The pathway interactions used by *PARADIGM* come from a variety of pathway sources, but they are primarily constructed from:

- National Cancer Institute Pathway Interaction Database
- BioCarta
- Reactome
- Kyoto Encyclopedia of Genes and Genomes
- Pathway Commons

These sources were selected because they consist of highly confident sets of richly annotated interactions that is a necessity for running *PARADIGM*. Unlike many pathway-based approaches, *PARADIGM* uses detailed information about how genes regulate each other, whether transcriptional, post-translational, through complex formation, or through a family of genes.

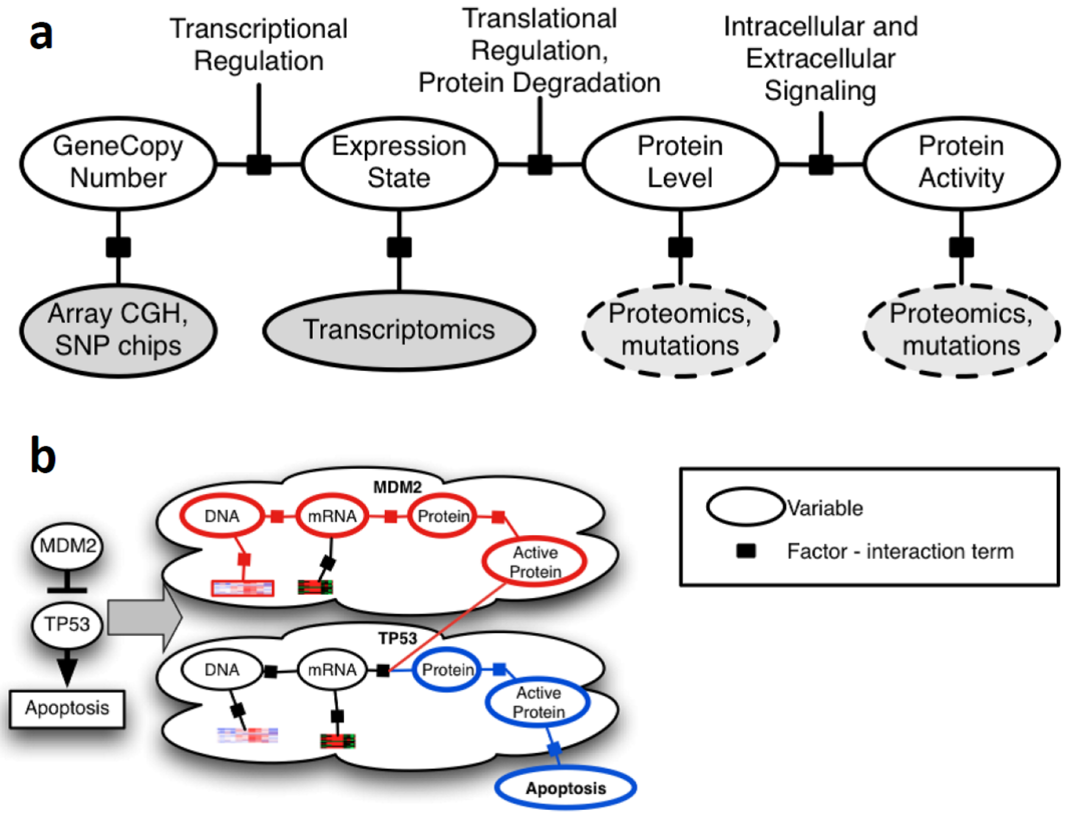


Figure 2 Gene dogma and interaction model for *PARADIGM* inference.

The modeling of these various sources of data and types of interactions is tied to the way the factor-graph model is represented. Figure 2 shows the two major components of the *PARADIGM* model, the gene model (a) and the interaction model (b). The gene model represents the various molecular states of a gene from whether or not it is present in the genome, transcriptionally expressed, to the expression and activity of the protein. Various data sources are then attached onto these nodes as evidence. The interaction model connects the active state of the parent node to the corresponding regulatory target of the child node based on the type of interaction in the pathways. In the example shown, the active state of *MDM2* is connected at the translational regulation factor for *TP53*.

Before supplying the data to *PARADIGM*, each dataset type is first transformed into rank-ratios. This is done by ranking all values across all samples from smallest to largest and then each rank *r* is transformed into the range *[0,1]* by the formula *(r-1)/(N*G-1)* where *N* is the number of samples and *G* is the number of genes measured. Briefly, *PARADIGM* then uses a belief-propagation algorithm on a factor graph derived from ϕ to combine gene expression, copy number, and genetic interactions to compute *inferred pathway levels* (IPLs) for each gene, complex, protein family, and cellular process. The IPL for a gene is a signed log-posterior odds of the state of the gene given the observed data. Positive IPLs reflect how much more likely the gene is active in the tumor, while negative IPLs reflect the negative log probability of how likely the gene is inactive in the tumor relative to normal.

In many cases the activity can be inferred from the direct and indirect regulatory influences that the gene participates. For example, *PARADIGM's* inferred activity of a transcription factor increases if several of its targets are overexpressed and amplifications in the genome cannot explain their up-regulation. By logical extension of this reasoning, *PARADIGM* may infer that a kinase is active if several of its targets have already been inferred to be active (e.g. the targets may be transcription factors with over-expressed targets).

1.4.1 PARADIGM Features Identify Trends Not Apparent with Gene Expression Data Alone

One of the first results by *PARADIGM* in the TCGA ovarian paper showcased the advantage of the factor-graph model by identifying the up-regulation of particular isoforms of *FOXM1* responsible for proliferation and DNA repair signaling. While, *FOXM1* was not clearly over-expressed in ovarian the expression of *FOXM1*'s downstream transcriptional target indicated high activity of *FOXM1*. This pathway was previously undescribed in ovarian cancer.

Also in TCGA ovarian, *PARADIGM* IPLs where shown to produce more accurate predictors of survival compared to gene expression signatures [Vaske 2010]. These results show that by incorporating the pathway model additional information is being captured by *PARADIGM* that is orthogonal to the information captured within the expression data alone.

1.4.2 Considerations when running PARADIGM

While *PARADIGM* is a powerful tool for analyzing genomic data, there are some caveats. Since genomic data is attached onto a pathway model, inferences cannot be made on genes that are not in the pathway models. This also means that incorrect or missing interactions will affect the inferences. This is not a major issue since many of the key pathways involved in cancer are relatively well studied, however, less studied pathways are poorly represented.

Another less obvious consideration is that message passing, the cross-talk between genes in the interaction model, will lead to correlations between nearby features that may be driven by a relatively small number of data points. Care must be taken to account for these correlations when drawing conclusions about *PARADIGM* results, otherwise results may appear much more significant than they are in reality. Throughout my work, I have incorporated statistical sampling to control for false positive conclusions resulting from such data modeling artifacts.

1.5 Summary

By incorporating pathways in with genomic data, statistical power is increased by drawing knowledge external to the dataset and guiding biological interpretation of results. In this chapter, I introduced the *PARADIGM* algorithm that lays the groundwork for my own research to be presented in this thesis. In the following chapters, I highlight my work with *PATHMARK* and *PARADIGM-SHIFT* for identifying pathway-based biomarkers that offer unique insight into the pathway differences across tumors and predicting the functional impact of mutations by observing the effect of mutation in pathway space. Lastly, I discuss my contribution to making my computational methods available for other researchers to run through the maintenance of the code available publically online and also usable as a Galaxy module.

2 PATHMARK

Molecular signatures derived from omics data can offer important insight into guiding treatment decisions or discovering new lines of therapy. They provide a window into the unique tumor of each patient; however, determining the truly significant features from a high dimensional dataset is challenging. At the gene level many of the alterations across a particular cancer cohort may appear random due to noise, but when a broader pathway view is taken a broader set of commonalities between different tumors can be identified. Determining which commonalities are significant is key for cancer treatment.

Certain pathway alterations are selected for within a tumor because they provide a survival advantage. Once an aberrant pathway is identified, researchers can test perturbations that disrupt these cancer pathways and hopefully disrupt tumor growth as well. *PATHMARK* aims to identify differential pathways in order to reveal molecular characteristics of the tumor, hopefully providing insight into how best to treat these tumors.

2.1 Subtype Specific and Independent Patterns Identified by PARADIGM Analysis in TCGA Colorectal

With *PARADIGM*, we can perform data integration across multiple data types and onto a pathway model. Studies mentioned in the previous chapter illustrated that *PARADIGM* improved survival prediction in ovarian cancers and revealed an uncharacterized activation of the FOXM1 transcription factor signaling pathway. In colorectal cancer (CRC), copy number, gene expression, methylation and pathway data were integrated using *PARADIGM*. The analysis showed a number of new characteristics of CRC. For example, despite the diversity in anatomical origin or mutation levels, nearly 100% of these tumours have changes in MYC transcriptional targets, both those promoted by and those inhibited by MYC, Figure 3. These findings are consistent with patterns deduced from genetic alterations and suggest an important role for MYC in CRC. The analysis also identified several gene networks altered across all tumour samples and those with differential alterations in hypermutated versus non-hypermutated samples.

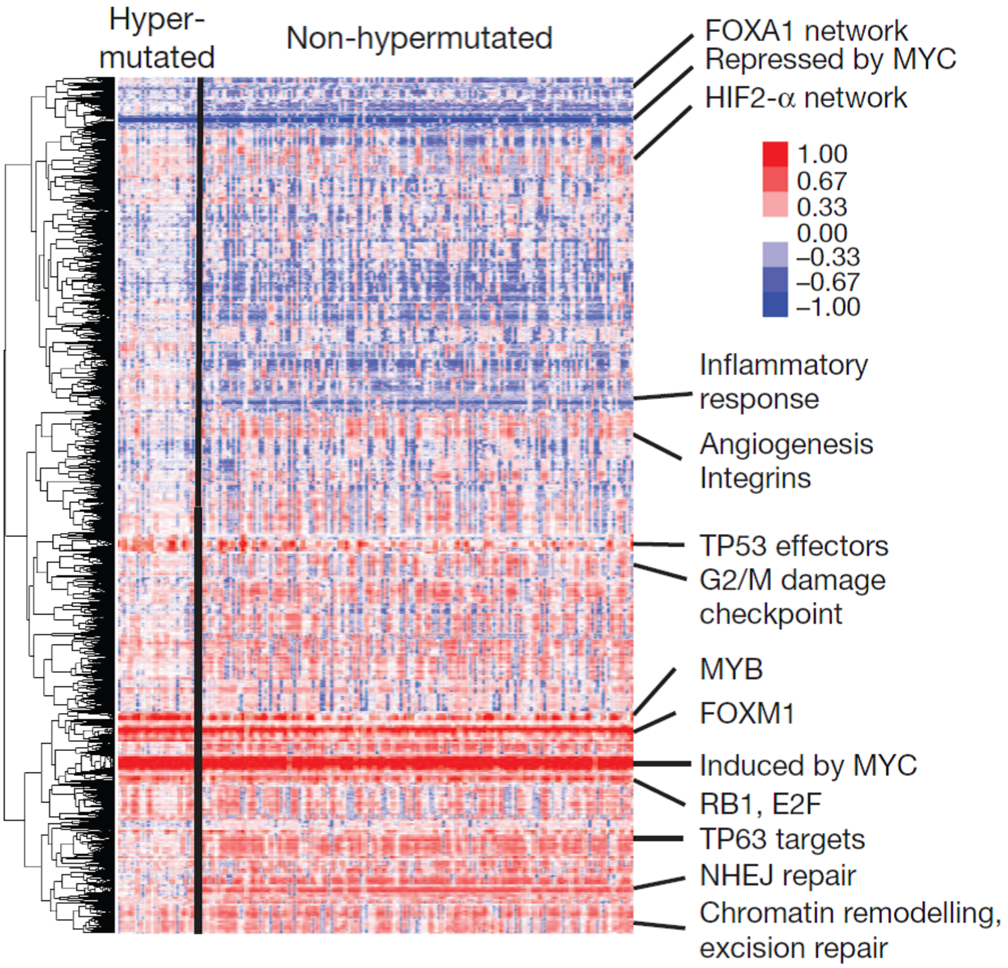


Figure 3 *PARADIGM* IPL heatmap for CRC.

Some of the trends on the heatmap for CRC are extremely striking, so identifying them is a simple task, however some of the smaller pathways are difficult to assess based on clustering. By utilizing pathway information to select pathway biomarkers, *PATHMARK* was developed to identify differential features that clustered in a region of the network. This would help guide annotation of important features outputted by *PARADIGM* since several features would often drive the selection for pathway biomarkers.

2.2 The PATHMARK Method

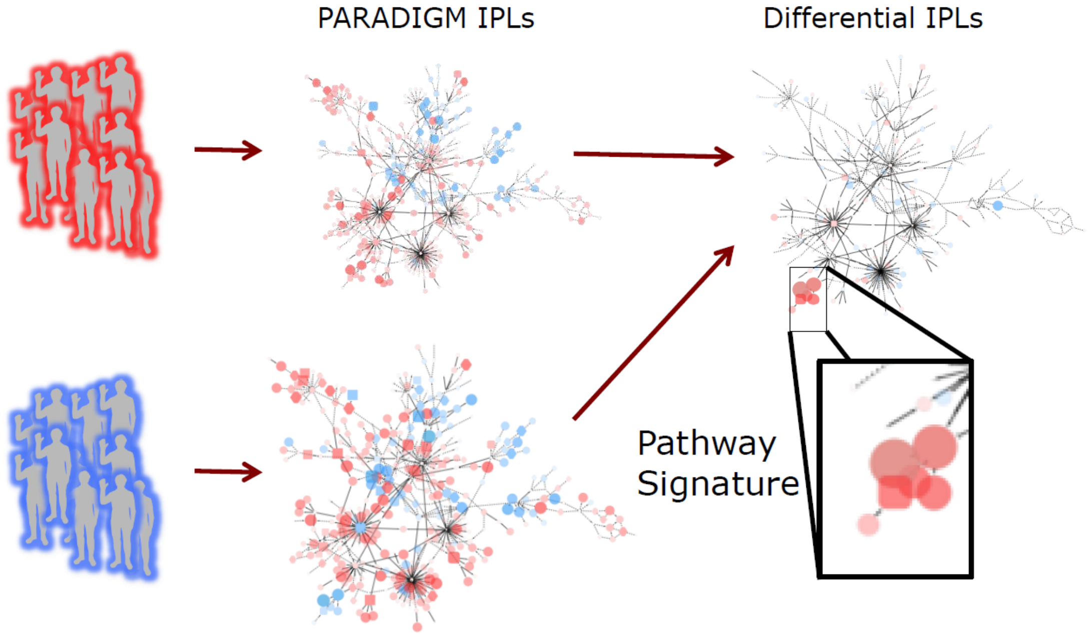


Figure 4 Overview of the *PATHMARK* methodology.

In order to identify significantly differential subnetworks from *PARADIGM*, *PATHMARK* was developed to select groups of pathway features that had high differential scores and were close in pathway space, Figure 4. This method was developed in conjunction with Ted Goldstein, Steve Benz, and Charles Vaske in order to provide a visual representation of the significant subnetworks generated by *PARADIGM*. Since *PARADIGM* IPLs are calculated for every pathway feature in the pathway, a selection of these features can be performed. First differential IPLs are calculated from the PARADIGM IPLs, similar to standard differential analysis, by comparing two groups of samples within a cohort. Such comparisons could include whether a sample is in a subtype versus not, or was resistant versus sensitive to drug treatment. A few different differential analyses have been implemented, such as significance analysis of microarrays (SAM), linear models for microarray data (LIMMA), and *t*-test. SAM is used most frequently since it has a correction factor for genes that have high differential with small variance, which can lead to over-inflated differential scores. LIMMA was updated to handle RNAseq count data, so it is also a popular choice when that data is available [Ritchie 2015]. Once differential scores are calculated for each *PARADIGM* pathway feature, *PATHMARK* performs a selection on the entire pathway to select subnetworks. This is done by keeping any edges that have source and target nodes with differential scores that exceed some threshold. This threshold is set by two parameters, the filtering parameters, which define the number of standard deviations above the mean value across all differential scores that the nodes have to exceed in order to be included. For example, parameters of 0 and 0.5 would mean that at least one of the two nodes would need to exceed 0.5 times the standard deviation over the mean, while the other node just needs to be above the mean. Typical parameter settings include (0.0, 0.0) and (1.0, 1.0) depending on the size of the resultant subnetwork desired. Since this selection considers pairwise nodes that have an interaction, *PATHMARK* typically enriches for regions with multiple data points supporting the conclusion that the pathway is differentially altered.

Two additional filtering steps are usually taken before viewing the final result to further refine the subnetwork result. The first filtering step removes “complex” pathway entities that do not have at least half of their gene components. This filtering is performed because “complex” pathway entities derive their entire IPL from the surrounding network, since there is no data directly associated with a “complex.” This potentially leads to long chains of these “complexes” that do not reveal anything meaningful since the values are driven by a single differential gene. Important genes could also be missed by *PATHMARK*, since a hard threshold is used for selection. To account for this another filter is applied to pull in any regulators that have at least half of their targets in the subnetwork (with at least 4 targets). This helps keep regions of the subnetwork connected, when hub genes do not make the cutoff.

To assess the significance of the *PATHMARK* network, the size of the real network is compared to that of a background distribution of *PATHMARK* solutions with permuted data. Since *PARADIGM* naturally propagates signal through the network, it would not be a fair comparison if this structure was broken in the *PATHMARK* nulls. Therefore, the permutation process is performed prior to running *PARADIGM* in which gene labels are permuted and then the null cohorts are run through *PATHMARK*. This asks the question of how likely a subnetwork of the observed networks size could have been observed if the genes interact randomly with each other with a similar network structure. The size of the network can be defined as the number of nodes in the entire network, the number of edges in the entire network, the number of nodes in the largest component, or the number of edges in the largest component. The number of nodes in the largest component is sufficient for determining significance and is typically the one used as it is the most intuitive when viewing the main component of the *PATHMARK* result.

Recently and in addition to significance assessment, robustness analysis was included to allow us to determine how robust regions of the *PATHMARK* solution are. This is performed by performing a bootstrap analysis in which bootstrap cohorts are created by sampling cohorts of similar size, but allowing for resampling (obtaining the same sample twice). The concept is that our data set is the closest set we have to the true population, so by resampling we look at the stability of certain regions of the complete *PATHMARK* solution. Nodes and edges can then be annotated with the proportion of times they were included in the bootstrapped *PATHMARK* solution to estimate the robustness of the subnetwork selection.

2.3 Pathway Signatures Identify Familiar Networks Related to Breast Cancer Subtype

We used the network analysis tool *PARADIGM* and newly developed *PATHMARK* to identify pathway-based mechanisms that underlie subtype-specific responses. *PARADIGM* was used with copy number and transcription data to calculate integrated pathway levels (IPLs) for 1441 curated signal transduction, transcriptional, and metabolic pathways. We compared IPLs for cell-lines and primary breast tumors using data from The Cancer Genome Atlas (TCGA) project and found a general concordance between transcriptional subtype and pathway activity across the two cohorts. This subtype-specific pathway activity likely explains much of the observed subtype specific responses. Mechanistic interpretation of IPLs for 1441 pathways is complicated by the overlapping elements in many of the curated pathways. We overcame this complication by merging the 1441 curated pathways into a single “SuperPathway” in which redundant pathway elements are eliminated. This approach enabled us to identify SuperPathway subnets that differed in activity between transcriptional subtypes through a newly developed analysis of *PARADIGM* IPLs called *PATHMARK*. *PATHMARK* identifies regions of the network, or subgraphs, with differential IPLs connected in cluster.

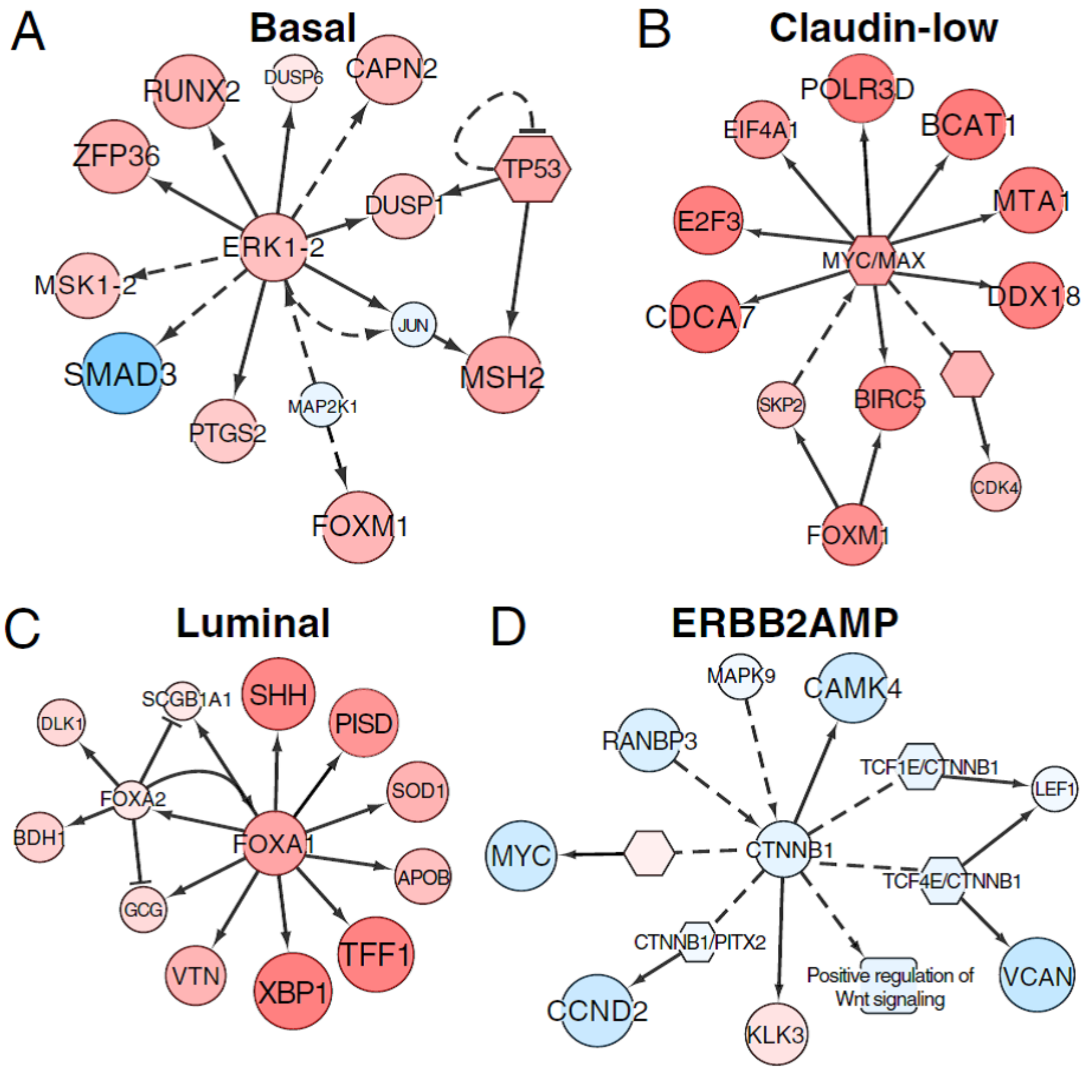


Figure 5 *PATHMARK* subnet heatmaps for four different breast cancer cell line subtypes.

As an example, comparison of subnet activities between basal cell lines and all others in the collection identified a network comprised of 1104 nodes (e.g., proteins, protein complexes, or cellular processes) connected by 1,242 edges (e.g., protein– protein interactions) between these elements. Several subnetworks were up- or down-regulated in the subtype-specific SuperPathway networks. Figure 5A, for example, shows up-regulation of an ERK1/2 subnetwork controlling cell cycle, adhesion, invasion, and macrophage activation. The forkhead box M1 and DNA-damage subnetworks also were up-regulated markedly in the basal cell lines. The claudin-low network showed up-regulation of many of the same subnetworks, as well as up-regulation of a MYC/Myc-associated factor X (MAX) subnetwork associated with metabolism, proliferation, angiogenesis, and oncogenesis, Figure 5B. Comparison of the luminal cell lines with all others showed down-regulation of an activating transcription factor 2 network, which inhibits tumorigenicity in melanoma, as well as upregulation of forkhead box A1 (FOXA1)/forkhead box A2 (FOXA2) networks that control transcription of estrogen receptor- regulated genes and are associated with good prognosis luminal breast cancers, Figure 5C. ERBB2AMP subnetworks were similar to those for luminal cells; this similarity is not surprising because most ERBB2AMP cells also can be classified as luminal. However, Figure 5D shows down-regulation of a β-catenin (CTNNB1) network in ERBB2AMP cell-lines; up-regulation of this network has been implicated in tumorigenesis and is associated with poor prognosis.

2.4 Differentiating Drug Response in Breast Cancer Cell Lines and Tumors

Surprisingly the *PATHMARK* analysis of differential drug responses among the cell lines also revealed subnet activities that provide information about mechanisms of response. The molecular subtypes of breast cancer are very distinct with differing prognosis and molecular features. Response to therapy is also highly correlated with subtype. Certain subnetworks within the subtype-specific *PATHMARK* results correctly predict response to therapy. For example, basal cell line sensitivity to the DNA-damaging agent cisplatin was associated with up-regulation of a DNA-damage response subnetwork that includes ataxia telangiectasia mutated and checkpoint kinase 1 homolog, key genes associated with response to cisplatin, Figure 6A. Likewise, ERBB2AMP cell line sensitivity to geldanamycin [an inhibitor of heat-shock protein 90 (HSP90)] was associated with up-regulation of an ERBB2-HSP90 subnetwork, Figure 6B. This observation is consistent with the known ERBB2 degradation induced by geldanamycin binding. On the other hand, AURKB pathway down-regulation in ERBB2AMP cell-lines indicated resistance to treatment with VX-680 [an inhibitor AURK], Figure 6C.

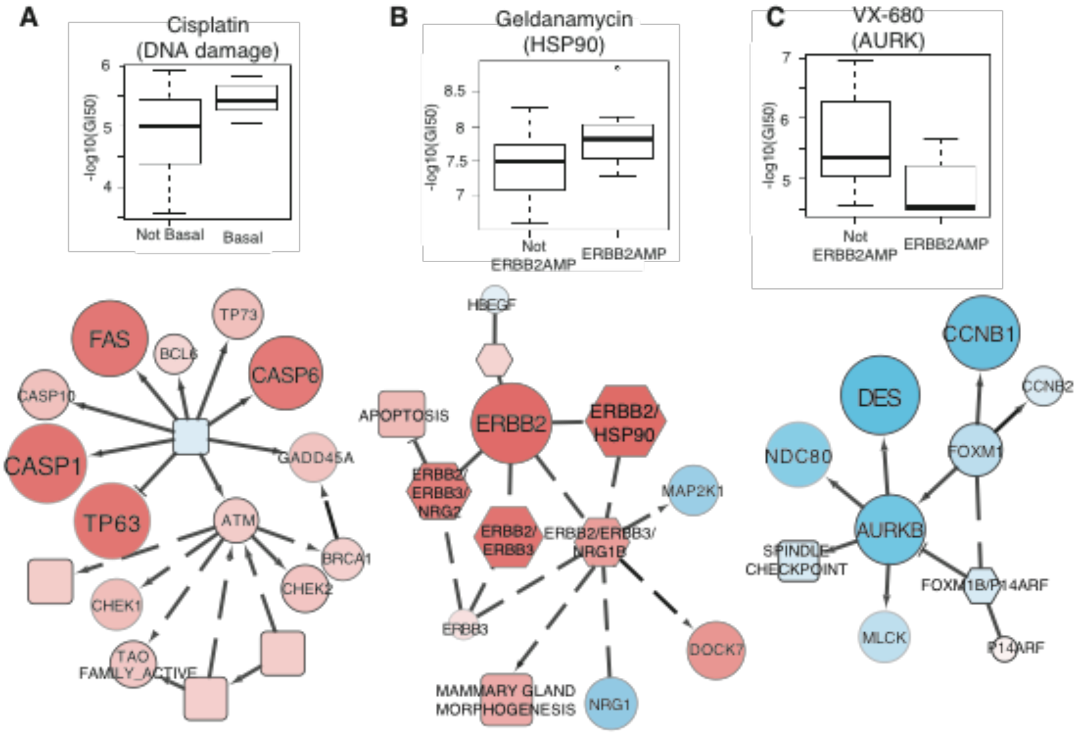


Figure 6 *PATHMARK* subnetworks are predictive of drug response in breast cancer cell-lines.

In the article *Whole-Genome Analysis Informs Breast Cancer Response to Aromatase Inhibition*, a similar finding was noted in breast cancer patients undergoing neoadjuvant aromatase inhibition. The *PARADIGM*-inferred pathway signatures were further used to derive a map of the genetic mechanisms that may underlie treatment response. In addition, PathScan was used to assess significant associations between mutations and pathways with aromatase inhibitor response. *PATHMARK* pathway biomarkers selected with high association with Ki67 biomarker status were consistent with the PathScan results, and among the largest of the hubs in the identified network were a central DNA damage hub with the second highest connectivity (55 regulatory interactions; 1% of the network) and TP53 with the 14th highest connectivity (26 connections; 0.5% of the network). Additional highly connected hubs identified in order of connectivity were MYC with 79 connections (1.4%), FYN with 45 (0.8%), MAPK3 with 43, JUN with 40, HDAC1 with 40, SHC1 with 39, and HIF1A/ ARNT complex with 39. As *PATHMARK* and *PARADIGM* do not integrate mutation data into the analysis, this approach provides orthogonal information to mutation analysis. Thus, *PATHMARK* which used genomic data and pathways without mutation data were able to identify the same significantly altered pathways relevant for aromatase inhibitor response.

2.5 Summary

As described in many of the results discussed in this chapter, *PATHMARK* is a useful tool for giving a broad overview of which pathways are altered in cancer cohorts. The networks obtained from *PATHMARK* often have features that are revealing of cancer disease mechanisms and of treatment response. However, from the PATHMARK result it is not clear which differentially activated or deactivated pathways are essential to the tumors survival (i.e. *drivers*), versus those that are a consequence of an upstream effect (i.e. *passengers*). To answer questions about what effects mutations might have on pathways and their connections, I developed *PARADIGM-SHIFT*, which will be discussed in further detail in the next chapter.

3 PARADIGM-SHIFT

The major mechanism by which cancer arises is through somatic mutations. These mutations can lead to alterations in gene regulation as well as changes in protein structure and function. Individual tumors can contain hundreds to thousands of mutations. It is critical to distinguish mutations that have an important role defining the cancer – *driver* mutations – from mutations that are unimportant to the tumor – *passenger* mutations. Differentiating *driver* and *passenger* events is essential for understanding cancer disease mechanisms, which can help guide treatment decisions as well as identify novel targets for treatment. Genomic probing with technologies such as expression arrays and high-throughput RNA sequencing provide insight into changes in gene regulation in cancer, but determining the tumorigenic role of a coding mutation is less clear. Genomic data coupled with pathway information provides insight into the functional impact of a mutation to particular genes. In this section, I introduce my method *PARADIGM-SHIFT*, and discuss how it is used to predict the functional impact of mutation – gain-of-function (GOF) or loss-of-function (LOF) that provides additional mechanistic understanding to *driver* mutations within certain cancer tumors.

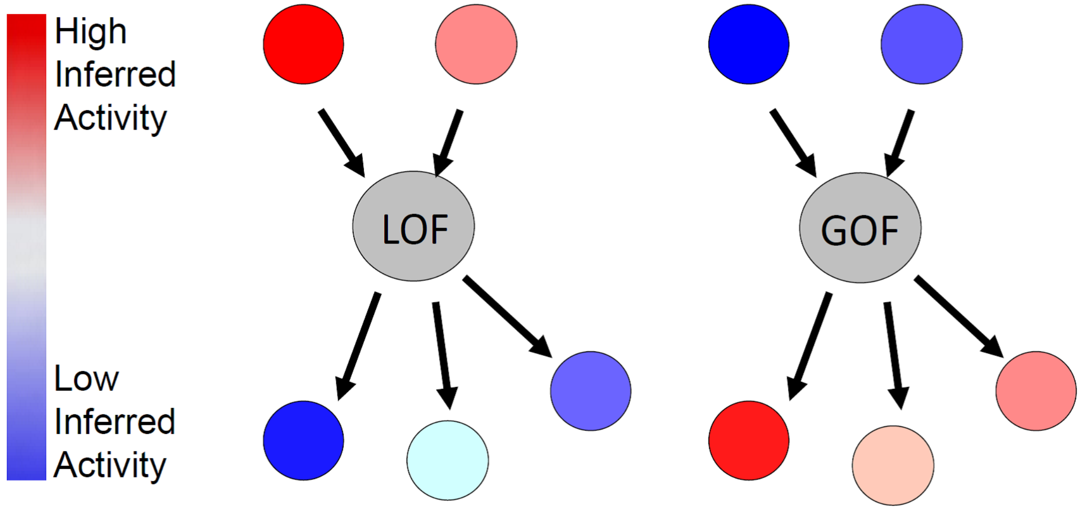


Figure 7 Anticipated pathway influences for gain- and loss-of-function mutations.

PARADIGM-SHIFT is a mutation prediction method based on integrated pathway analysis to discriminate LOF, neutral, and GOF mutations. Utilizing the set of regulatory interactions annotated for a given gene, it can detect a discrepancy in the downstream effects of an altered gene compared to what is expected from its upstream influences, Figure 7. Intuitively, if a mutation causes a LOF to a focus gene (FG) then it may create a particular signature on that FG's pathway neighborhood. In the case of a LOF event, the regulatory input to the FG would indicate that the gene should be turned on at the transcriptional and/or post-transcriptional levels. For instance, a transcription factor and kinase that regulate different parts of the FG's activity may themselves be active given the data on a particular sample. However, when one inspects the sample's data for neighboring genes downstream of the FG one would find evidence that the FG is not active. For example, FG itself may be a transcription factor that activates several target genes all of which have low expression levels in the sample. In the GOF case, the opposite situation would occur where the downstream targets are consistent with a higher activity of the FG than what would be expected from the FG's regulatory inputs.

Feedback circuitry in the cell may accentuate the difference in upstream and downstream activity information. Cells may detect that the function of the FG is lost and signal through feedback circuitry in an attempt to "rescue" FG by activating its upstream regulatory inputs. When this happens, the upstream signals can become even more strongly suggestive of high levels of FG's activity even though no evidence of that activity is present downstream. A method that leverages such discrepant pathway information surrounding the FG may have a chance of predicting the consequences of a mutational event and distinguish cases in which they are neutral, loss-of-function, or gain-of-function. In this chapter, I will show that a score based on this discrepancy is highly predictive of the presence of a mutation and that the directionality of this discrepancy also reflects the gain- or loss-of-function in a gene.

This additional information could be essential in guiding treatment decisions and determining additional therapeutic targets in the future. In some cases genes are known to have cases of both gain- and loss-of-function mutations, such as in *NOTCH1* discussed later in this section; in such a situation it is possible that a targeted therapy given to counteract a suspected gain-of-function mutation would actually lead to worse outcome in a patient whose tumor mutation actually produced loss-of-function. Understanding the mechanistic consequence of a mutation is as important as identifying *driver* mutations. Further discovery of new gain-of-functions could also lead to the development of new targeted therapies.

Application of my method to a set of known driver mutations reveals that there is a significantly strong signal for loss- and gain- of functional mutations in the surrounding network, demonstrating the sensitivity of this approach. In addition, when applied to the negative control of passenger mutations, the method predicts little pathway impact, indicating this approach also has high specificity.

A comprehensive cancer survey such as that being generated by TCGA uncovers numerous genomic events that are a mix of both causal, *driver* events and incidental *passenger* events that accumulate as a result of dysregulated genomic surveillance and cell proliferation with clonal expansion over time. Exome and whole-genome sequencing uncover recurrent mutational events in a few genes and many low frequency events in many others. Importantly, many of the low frequency genes are known to be functionally important in the tumors in which they arise. For example, while *BRAF*$^{\mathrm{V600E}}$ is common in melanoma, it occurs in only 3% of non-small cell lung cancer, but is clearly a driver when present in both of these tumor types [Dankort 2007]. Of the many significantly recurrent mutations in various cancer types most are not well understood, hopefully pathway-based methods like *PARADIGM-SHIFT* can shed light on some of these.

3.1 Orthogonal Methods for Assessing Mutations

Several computational methods for predicting the functional importance of mutations exist. These methods often use the frequency of a mutated gene across a cohort, the location of a mutation in the gene, whether the mutations are silent or non-silent, frame-shifting, potentially protein domain altering, found in more evolutionarily conserved regions of the peptide sequence, or cluster together in the protein sequence or structure. While these existing methods have shown tremendous success, they each have certain limitations that impact their generality. For example, some methods must be trained from external datasets such as from the COSMIC database that introduces possible circularity to the analysis and biases the discovery of genes whose mutational impact has already been characterized.

SIFT [Ng 2003] and *MutationAssessor* [Reva 2011] classify the functional impact of mutations based on sequence conservation at the positions in which these missense mutations occur. If a particular position were highly preserved across many normal genomes, a mutation there would be predicted to be highly impactful. This methodology functions upon the premise that highly conserved regions are important for normal protein function, so that a disruption is negatively selected for in the normal setting. Since *driver* mutations are likely to result in a significant change in function these methods take advantage of the fact that these mutations are more likely to fall in a conserved region to account for that large change in function. *PolyPhen-2* [Adzhubei 2010] similarly predicts the functional impact of individual events, but includes additional biological predictions external to *SIFT*. By taking into consideration features related to the local sequence of the mutation and structural information trained against a library of known damaging alleles of human disease, Polyphen-2 was able to improve performance over SIFT.

The genomic landscape undergoes a plethora of alterations on the path to carcinogenesis from copy number gains and losses to mutations. In order to make sense of the molecular mechanisms of cancer it is essential to distinguish the *driver* events from a “sea” of *passenger* events. Since the impact of missense mutations is less obvious to predict, many approaches have been developed to distinguish *driver* or *passenger* mutations. Because of positive selection pressure, we can identify *driver* mutations as occurring at a higher frequency than expected by chance or based on biological information such as predicted impact.

Methods such as *MutSig* [Getz 2007] take advantage of positive selection pressure within tumors to detect *driver* events. Genes are scored by comparing the frequency of mutations to a *passenger* mutation rate estimated by excluding known *driver* mutations. Special considerations are also taken within the null model to handle genes of different size, genes of larger size are more likely to have a mutation occurring at random, and differences in mutation rate due to locations in the genome, based on GC content or proximity to replication fork machinery. *Passenger* mutation rate is estimated by looking at mutations in non-coding gene regions, however this estimation is not perfect since there are non-coding genes as well as regulatory mutations that can also be positively selected for during carcinogenesis. The significantly mutated genes are determined by setting a threshold based on the false discovery rate.

By combining the advantages of recurrence based methods and functional impact based methods, *OncodriveFM* [Gonzalez-Perez 2012] is able to filter out some of the more recurrent *passenger* mutations that have low predicted functional impact and retain lowly recurrent *driver* mutations with a high functional impact. Many *passenger* mutations are often called significant because of various biases towards mutation, such as proximity to the replication fork, while rarer *driver* mutations are missed because of a lack of statistical power in recurrence methods. *OncodriveFM* weights each mutation based on their functional impact score to create a combined score based on functional impact and recurrence.

There are also methods that take advantage of prior knowledge about related genes in the form of pathways to predict *driver* mutations. *MEMo* [Ciriello 2012] and *Dendrix* [Vandin 2011] leverage pathway information to call *drivers* by identifying subnetworks or related gene-sets with genomic events that are highly mutually exclusive. The idea is that alterations in a single gene in a pathway is sufficient to perturb it, therefore, mutual exclusivity is often observed for genes affecting the same pathway since once one hit is achieved a second is not necessary for the cancer. *HotNet* [Vandin 2010] similarly identifies a subnetwork of mutations, but by identifying clusters of mutated genes through a heat diffusion algorithm.

Gene signature-based approaches train machine-learning classifiers to recognize the presence and absence of mutations from molecular features such as gene expression data [Mooney 2011]. These methods can be applied to any number of genomic perturbations including mutations, focal copy number gains or losses, or methylated promoters. They can be applied to a variety of both coding and non-coding mutations and thus are potentially capable of detecting whether mutations in regulatory regions have functional significance. Gene signatures are first computed from a training set of data in which either a subset of data or an external dataset is used. Genes with expression levels that are differentially associated with the presence (compared to the absence) of a mutation are candidates for inclusion in the gene signature using any number of a variety of univariate and multivariate machine-learning and feature selection approaches. One major obstacle in this work is the identification of signatures that truly are robust enough to generalize from one dataset to another. Researchers have faced the difficulty of combining various microarray-based and now high-throughput sequencing based platforms together as well as the inherent stochastic nature of gene expression.

To date no existing method makes use of genetic pathways to interpret the functional consequences of a mutational event. However, the availability of multi-dimensional datasets for cancer samples like those generated by the TCGA project make such an endeavor possible. If our pathway knowledge surrounding a particular gene is complete enough and we have enough data to provide information about the activity of neighboring genes, then we can use that knowledge to measure the pathway consequences of a mutation.

3.2 The PARADIGM-SHIFT Method

PARADIGM-SHIFT [Ng 2012] takes advantage of many of the same principles used in the methods described above to predict not only *driver* from *passenger* events, but also to predict whether mutations are likely to increase the function, gain-of-function (GOF), or decrease the function, loss-of-function (LOF), of the gene. This is achieved by estimating the functional impact of a mutation on the pathway by using *PARADIGM* [Vaske 2010]. In cases of LOF or GOF mutation, there would be conflicting signal from the genomic data upstream of a mutation and downstream of a mutation. For example, in the case of a LOF the upstream regulators may be trying to turn the gene back on, but since the missense mutation renders the protein non-functional the downstream targets are not active. These impact predictions, or shift scores, can be calculated for each sample regardless of mutation status and a LOF or GOF call is determined if there is a high recurrence of low or high shift scores in the mutated samples in comparison to the non-mutated samples. Since genomic data and pathway data is utilized instead of sequencing data, *PARADIGM-SHIFT* provides an orthogonal view to traditional mutation analyses.

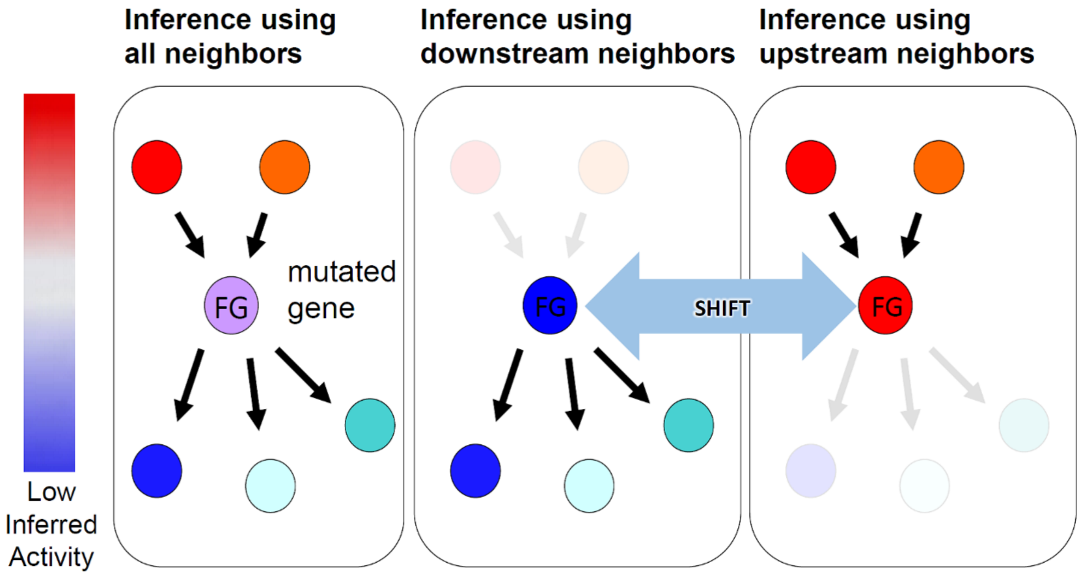

Inference using all neighbors
Inference using downstream neighbors
Inference using upstream neighbors
mutated gene
FG
FG
FG
SHIFT
Low Inferred Activity

Figure 8 Discrepancies in pathway signals can be isolated through *PARADIGM*.

The core of our approach estimates the shift in each tumor sample for each focus gene (FG) using two runs of the original PARADIGM algorithm, Figure 8. In the downstream run, the FG is left connected to a neighborhood of its downstream targets while upstream regulators are disconnected. In the upstream run, a neighborhood of upstream regulators is left connected to FG but all downstream targets are disconnected. As with *PARADIGM*, all variables are trinary representing whether a feature is more active in the tumor relative to normal, more inactive in the tumor than normal, or the same level in tumor as in normal. The shift score then computes the difference between the inferred activities of FG determined in the downstream run from those determined in the upstream run.

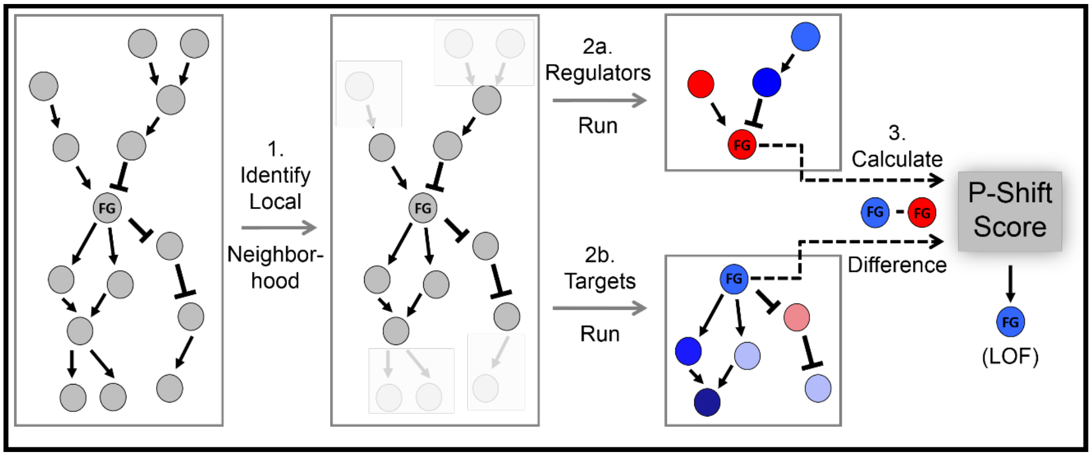


Figure 9 Detailed diagram of the *PARADIGM-SHIFT* method.

Since *PARADIGM-SHIFT* relies on network information in order to call significant impacts, sufficient pathway knowledge is required in order to identify a result, where there may be one. Most of our networks are derived from large pathway databases as described in the first chapter, but in some cases smaller constituent pathways of interest are pulled in for prediction on certain mutations. In addition to genomic data that *PARADIGM* needs in order to execute and a set of *PARADIGM* parameters trained on the network and genomic data, *PARADIGM-SHIFT* requires mutation data that contains information on which samples have mutations in any genes of interest. Figure 9 illustrates the steps *PARADIGM-SHIFT* takes in order to calculate the shift scores for each sample. First, feature selection is employed to determine the set of neighboring upstream and downstream features in the model. Typically, features are selected by including features based *t*-statistics in the top 84th percentile, or approximately above a standard deviation over the mean. The *t*-statistics are calculated using a standard *t*-test comparing two samples with unequal variance and size on the expression ensuring that the underlying data of the features selected in the neighboring networks differentiate alteration status. Once the neighboring networks have been selected the inferred activity of the upstream and downstream can be determined through two *PARADIGM* runs; one where only the connections with the upstream regulators are retained (R-run) and one where only the connections with the downstream targets are retained (T-run). Finally, the shift score is then calculated as the difference between the inferred activity of the downstream run and the inferred activity of the upstream run.

The accuracy of the trained model can be assessed by using the absolute shift score as a classifier to predict the presence of mutation in a cross-validation setting. In the case in which a mutation has a functional impact on the pathway, I would expect to observe high absolute shift scores, higher discrepancy between upstream and downstream signal, in mutant compared to non-mutant samples. If the model is predictive of a functional impact, then *PARADIGM-SHIFT* then makes a prediction of LOF or GOF based on the distribution of shift scores for the mutated and non-mutated samples, negative shifts for LOF and positive shifts for GOF. The strength of the predicted impact, which I define as the *mutant-separation*, is quantified using a *t*-statistic computed from the distribution of shift scores for the mutant contrasted against the non-mutant samples. The significance of the *mutant-separation* is computed by comparing the observed to a background model determined using the same fixed network model, but where gene labels are permuted for the input genomic data. A rigorous description of the procedure is given in the next section.

3.2.1 Detailed Description of the PARADIGM-SHIFT Score

We derive a *Pathway Shift* (*PS*) score based on the intuition of comparing the observed downstream consequences of a gene's activity to what is expected from its regulatory inputs. The *PS* score has the form:

$$PS(f) = \log\left(\frac{Observed(f)}{Expected(f)}\right), \qquad (1)$$

where the observed activity for *f* is derived from the downstream targets and the expected activity for *f* is derived from the upstream regulators. The caveat of course is that we never get to truly "observed" gene *f*'s activity so we must infer it from the activity of downstream targets. As implied above, the estimation of such an activity is necessarily recursive, requiring us to first estimate the activity of the downstream targets for *f* before we can predict *f*'s own activity. The computation is naturally framed as an inference problem over a set of interdependent variables, some of which are hidden. In the last couple of decades, efficient procedures have been developed to compute the probabilities of a system of variables connected together in a probabilistic graphical model [Friedman 2004].

To estimate pathway-neighborhood dependent inferences on focus gene *f*'s activity, we restrict our view of the data to subsets of features in ϕ. If $R \subset \phi$ is the set of regulators of *f* and $T \subset \phi$ is the set of targets, then let *D(R), D(T),* and *D(f)* refer to the data observed for the regulators, targets, and the focus gene *f* respectively. Likewise, the interactions are restricted to the subset of features in the focus gene's neighborhood and denoted ϕ(T) to represent the pathway features and interactions involving only the targets to each other and to *f* itself. Similarly, $\phi(R)$ represents the same for the upstream regulators. With these definitions in hand we write the *PS* score as the following log ratio of two constituent likelihood-ratios:

$$PS(f) = \log\left(\frac{LR\left(D(T) \mid x_f^a, \phi(T)\right)}{LR\left(D(R), D(f) \mid x_f^a, \phi(R)\right)} \right), \tag{2}$$

where $LR(Y|x^a,Z)$ is defined as $P(Y|x^a,Z)/P(Y|x^{\neg a},Z)$, the likelihood ratio computed over one possible alternative value for X, x^a, compared to the probability of the other two possible values – less active in tumor x^i and similarly active in tumor x^0 – the combined event $x^{\neg a} = \{x^i, x^0\}$ is written for short. Note that only the expected term in the denominator contains the entry *D(f)*, which represents the actual data for the focus gene of interest. This reflects the assumption that the data on the focus gene provides evidence for the *cis*-regulatory state of the gene and so is included among the regulators for *f*. Note that if data on the direct activity for *f* were instead available, such as phosphorylation status or enzymatic activity, then that data could be considered for inclusion into the numerator term for the observed targets. The quantity in equation (2) reflects the degree to which the observed data for the targets is consistent with high activity of the focus gene relative to the observed data for the regulators and the gene in question. There was a typo of $x_f^{\neg a}$ in the denominator in the original publication, which has now been corrected.

Further expansion of the *PS* score reveals the method by which it can be computed using the original PARADIGM algorithm. Application of Bayes Rule gives:

$$PS(f) = \log\left(\frac{P\left(D(T), x_f^a \mid \phi(T)\right)}{P\left(D(T), x_f^{\neg a} \mid \phi(T)\right)}\right) - \log\left(\frac{P\left(D(R), D(f), x_f^a \mid \phi(R)\right)}{P\left(D(R), D(f), x_f^{\neg a} \mid \phi(R)\right)}\right) - prior \quad , \qquad (3)$$

where *prior* is the log-prior-odds and has the same form as the first two terms in the equation except that all entries involving *D* are dropped. Another application of Bayes Rule would show that the first two joint probability ratio terms are equivalent to the LPO that the gene is active given either the state of the downstream targets (left-hand term) or the upstream regulators (right-hand term). The advantage of writing the joint probabilities in this form shows explicitly those terms of the form

$P(D, x \mid \phi)$ that are each efficiently computed with a message-passing belief propagation procedure on the underlying factor graph encoded by ϕ. The message-passing procedure takes care of summing out all of the hidden variables present in ϕ – the states of complexes, cellular processes, and the activities of all other genes other than *f*. The computation implements an iterative form of the Expectation-Maximization procedure that sequentially updates all variables by forming a running average until either a convergence tolerance of 10^{-9} is reached or 10,000 maximum iterations are exceeded. The code is freely available through the libDAI C++ open source library [Mooij 2009]. In the R-run version of PARADIGM, the LPO shown in Equation (3) and its corresponding log-prior odds are computed in two separate full factor graph convergence runs. Likewise, the T-run involves two separate EM runs to compute its two terms in Equation (3). Thus, in total, the computation time involved to compute the *PS* requires four EM convergence runs, but each task is run on a reduced pathway representation involving only the neighborhood of the focus gene. Thus, the computation time to calculate a *PS* for an entire dataset requires *2k* PARADIGM runs where *k* is the number of mutated genes in the cohort.

In practice we use the inferred pathway levels (IPLs) from PARADIGM for the computation of the *PS*. Specifically, we set $PS(f) = IPL_T(f)\text{-}IPL_R(f)$, where $IPL_{T|R}(f)$ is the IPL derived from the T- or R-run. The IPL is a signed LPO that always puts the highest probability state for *f* in the numerator. If the inactive state is in the numerator the IPL gives the negation of the LPO. This quantity is similar to Equation (3) except that the highest probability states determined in each run are contrasted. In the case where the active form of the gene is the most probable in each case the two formulas are equivalent. Finally, we found that a transformation of the *PS* score to a Z-score provided better overall results. Each gene's local neighborhood could have a certain bias to lean toward either positive or negative scores. To account for this we constructed 100 random samples for each gene by shuffling data tuples around the SuperPathway. This effectively associated random data with each gene's neighborhood. *PS* scores were calculated for each of these 100 samples and each *PS* was then normalized by subtracting the mean and dividing by the standard deviation determined from this simulation.

For computational efficiency, we use a local neighborhood around each gene rather than the entire network. Using the local neighborhood provides a good approximation of the full network as genes far away are expected to exert a diminishing influence on the inference of the focus gene. We tested including neighbors at distances 1, 2, and 3 and full for the positive controls. The empirical results indicate that the distinction degree increases only moderately after distances of two. To build the neighborhood, we traverse the graph and include any pathway features when there is at most one other intervening protein between the feature and FG. All interactions between the selected features were included in the neighborhood. If a protein was present in both the upstream and downstream neighborhoods, due to feedback circuitry, it was excluded from both the R- and T-runs.

3.2.2 Initial Findings on Published TCGA Datasets

We downloaded gene expression, copy number, and exome-capture mutation data for patient tumor samples from TCGA data coordinating center for 185 glioblastoma multiforme (GBM) samples on 7/28/12, 354 ovarian serous adenocarcinoma (OVCA) samples on 7/28/12, 219 colorectal carcinoma (CRC) samples on 7/28/12, 184 lung squamous carcinoma (LUSC) samples on 10/26/12, and 525 breast carcinoma (BRCA) samples on 9/24/12. Datasets for each tissue-specific tumor type were used separately as the dataset *D* for inferring mutation *PARADIGM-SHIFT* impact. We formed a comprehensive cellular pathway diagram for ϕ by merging together several pathway sources including NCI-PID [Schaefer 2009], Reactome [Matthews 2009], and BioCarta [Nishimura 2001] and then combining them into a superimposed pathway henceforth referred to as the "SuperPathway." We have previously described the construction of the SuperPathway for application to the analysis of a set of breast cancer cell lines and their response to various therapeutic agents [Heiser, Sadanandam et al. 2011].

We applied *PARADIGM-SHIFT* to a set of three well-characterized genes including RB1, TP53, and NFE2L2. The retinoblastoma (RB1) gene is a well-known tumor suppressor gene and plays a crucial role in the control of the G1→S transition of the cell cycle. We applied our method to predict the functional consequences of mutations to RB1 in the GBM cohort. Neighborhood selection identified six proteins in the upstream neighborhood of RB1 and ten downstream targets. Of these, 6 out of

9 samples received negative *PS* scores consistent with RB1's characterized role as a tumor suppressor.

We asked whether the distribution of *PS* scores were significant by comparing the *PS* score associated with the mutated samples to the *PS* scores associated to those samples without reported RB1 mutations. It is important to note that our method makes no use of the mutation calls in any way when deriving the score. While most of the samples received negative *PS* scores, there were a significant proportion of the mutated samples with shifts near the mean level seen in the non-mutated cases. These may reflect the set of samples harboring neutral passenger mutations that happen to land in the RB1 gene. It would be interesting to compare the clinical outcome of these RB1 mutant cases to those with low *PS* scores to see if their tumors are less aggressive.

While a *t*-statistic indicates that the distributions of the *PS* scores are appreciably lower for the mutants compared to the non-mutants, we performed a permutation analysis to assess whether the observed *t*-statistic was significant using a non-parametric approach. We formed random neighborhoods for RB1 by assigning data tuples from random genes to the regulators and targets of RB1. Using 1000 different sets of randomly assigned neighbors the entire procedure was repeated and the difference between the mutant and non-mutant distributions were computed. This test indeed revealed that the lower *PS* scores observed for the mutant RB1 samples were significantly lower than the non-mutants relative to those differences seen in these random controls.

TP53 is the most commonly mutated gene in cancer. It is a tumor suppressor and in most cases exerts a dominant negative action over the wild-type allele. In other cases, deletion of the gene resulting from loss of heterozygosity (LOH) or double somatic events are observed as early and frequent events in cancers spanning many tissue types. We compared our algorithm's ability to predict the functional impact of TP53 mutations in GBM and OVCA both of which have been published by the TCGA consortium. In GBM nearly a third (48) of the samples have a mutation in TP53. Of these, 19 had negative *PS* scores, 9 had positive *PS* scores, and the rest had near-neutral as determined by permutation analysis. In OVCA, the majority (67%) of the samples had a reported TP53 mutation. Importantly, it is believed that nearly 100% of the samples harbor such a mutation even though less than 100% were detected (Consortium 2011). The difference in *PS* scores for mutated versus non-mutated were again found to be significantly left-shifted.

3.2.3 Concordance with Recurrence-based Methods

To gauge the general agreement with predicting functional impacts for mutated genes that are considered to be driver rather than neutral passenger events, we compared our approach to *MutSig* [Getz 2007]. *MutSig* considers the frequency of the mutation, the location of the mutation in the gene, and several other features to calculate a significance score relative to an estimated sample-specific background mutation rate. We collected all *MutSig* scores for those genes that had representation in the SuperPathway and that had at least three mutations in GBM. We used the absolute value of the average of the *PS* scores and dividing them into two groups according to whether they were associated with significant or insignificant *MutSig* scores, Figure 10A. The results show a clear enrichment for higher absolute *PS* scores (either indicative of GOF or LOF) for those genes with significant *MutSig* scores compared to those with insignificant scores.

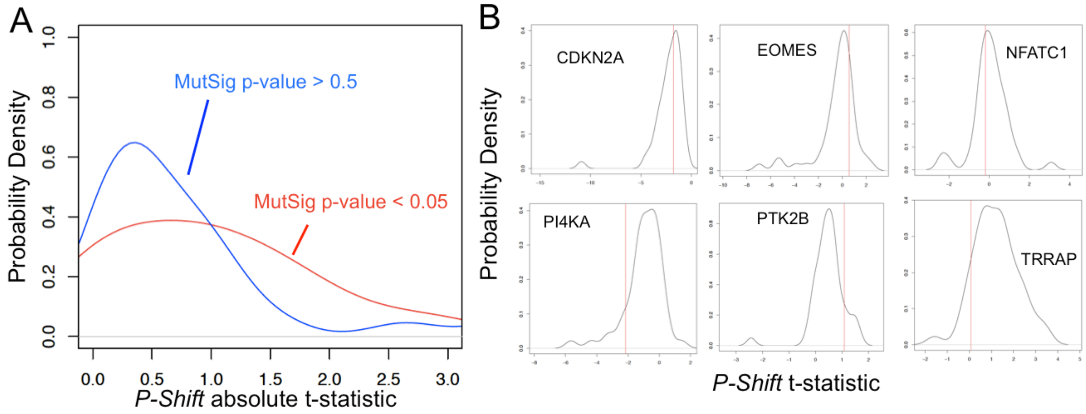


Figure 10 *PARADIGM-SHIFT* predictions concordant with *MutSig*.

To determine a rough estimate for the specificity of our approach, we collected six of the genes that received insignificant *MutSig* scores on which to perform the aforementioned permutation analysis. We plotted the *PS* scores from the permuted samples and found that in each of the six cases, the calculated *PS t*-statistics fell well within the range seen in the permuted controls, Figure 10B. We find that these mutations in genes with low *MutSig* scores are associated with *PS* scores that do not discriminate between mutant and non-mutant samples, consistent with the assumption that many of these mutations represent passenger events. Thus, our pathway-based method shows a degree of confirmation to a purely sequence-based analysis of mutational events.

As discussed in a previous section, most methods to assess significance of mutations in a given gene rely heavily on the prevalence of mutations across a clinical cohort with shared characteristics (e.g. early stage colon cancer). However, some rare events are of paramount importance to the patients in whom they occur, such as in the Ras-MAPK pathway in GBM. To determine novel impactful events, we applied *PARADIGM-SHIFT* to all of the mutated genes in GBM, OVCA, and LUSC. Our analysis identified probable gain-of-function mutations in MAPK1 based on only three out of 171 samples sequenced in GBM. All three lie in the protein kinase domain and two are predicted to change the kinase function due to their occurrence in highly conserved residues in the kinase. These results suggest kinase inhibitors targeting the ERK proteins in select cases may be effective. CDKN2A is a well-known tumor suppressor whose loss, primarily through homozygous copy number deletion, is an early driver of oncogenesis. Thus it is consistent that our method predicts LOF for this important tumor suppressor in both the GBM and LUSC for those cases in which the gene is present but mutated. Counter to expectation, our method assigns a positive score to NF1, which is a well-known tumor suppressor through its characterized inhibition of Ras. Detailed inspection of such examples may reveal important further refinements to the method particular in the neighborhood selection step.

In addition to NFE2L2 discussed earlier, analysis of the LUSC cohort also reveals potentially therapeutically important targets in select cases. PIK3C2G for example, may act as a driver in a handful of patients and could in theory be targeted by AKT pathway inhibitors or rapalogs. The low negative value of HUWE1, a less well-studied E3 ligase, suggests this enzyme might play a role analogous to that of CUL3 in degrading NFE2L2 in some cases.

PARADIGM-SHIFT applied to the ovarian dataset also gave informative insights into this tumor type. Nearly all samples harbor TP53 mutations (n=179). The *PS* was mostly negative for TP53 consistent with the expected LOF of this tumor suppressor. In addition, our analysis may clarify potentially important directional information about pathway alterations. For example, the EPH receptor family is known to participate in bidirectional signaling [Aoto and Chen 2007]. The high absolute differences in the *PS* scores seen across this diverse family in the ovarian cohort may reflect functionally opposing roles of these bidirectional receptors in oncogenesis.

3.2.4 Orthogonality with Sequence-based Methods

Finally, because *PARADIGM-SHIFT* is the first method to make use of a surrounding estimate of pathway activity to predict the impact of a mutation, we sought to measure the degree to which it provides orthogonal information compared to other popular approaches. For this comparison, we used SIFT [Kumar, Henikoff et al. 2009], PolyPhen2 [Adzhubei, Schmidt et al. 2010], and MutationAssessor [Reva, Antipin et al. 2007], each of which implements a different, sequence-based method to predict the consequence of mutations. We also included CONDEL [Gonzalez-Perez and Lopez-Bigas 2011], which produces an integrated call by combining the above three methods. We calculated the Pearson correlation between each of the methods and between each of the methods to *PARADIGM-SHIFT*. Not surprisingly, due to the heavy sequence-based nature of the previous methods, they all have higher correlations among themselves than they do to *PARADIGM-SHIFT*, Figure 11. Thus, our method may provide novel viewpoints on mutations that can be used in conjunction with sequence-based methods to gain a fuller understanding of the impact of mutated genes, their role in carcinogenesis, and how therapies might be developed for individual tumors.

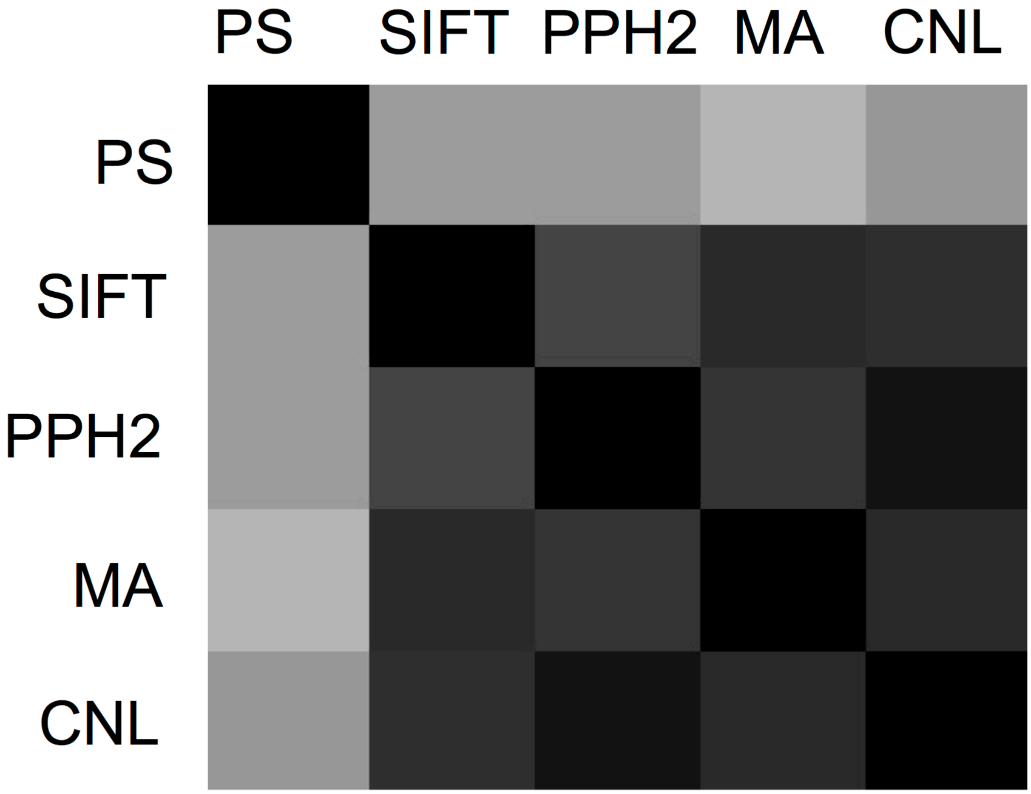


Figure 11 *PARADIGM-SHIFT* correlations to sequence-based approaches.

Our approach uses different information which may provide a complementary view compared to protein-sequence based approaches. It enables probing into infrequent events and can be used to detect the impact of non-coding mutations. In addition, it may be useful for detecting those cases that harbor passenger mutations where the mutation is either neutral or the cell has compensated somehow to keep the surrounding pathway intact. Finally, since our approach couples single gene mutation events with broader pathway activation signatures, it could be used to place genes with unknown/little known function and provocative mutations, into new pathways, as suggested by the case of HUWE1 above.

3.3 NFE2L2 in TCGA LUSC

To gauge the utility of the method in predicting gain-of-function mutations on a known proto-oncogene, we applied our method to mutations in NFE2L2 in lung squamous cell carcinoma (LUSC). NFE2L2 is a transcription factor that directs response to stress and oxidative damage in cells. Activating mutations in specific lysine residues stabilize the protein by preventing its degradation via binding to the KEAP1/CUL3 ubiquitin ligase complex. In the TCGA lung squamous dataset, NFE2L2 was found to be predicted as GOF by the enrichment of positive *PS* scores associated with NFE2L2 mutations compared to NFE2L2 wild-type samples. There are several features upstream and downstream of NFE2L2 that can explain the discrepancy in signal we are observing with *PARADIGM-SHIFT*. KEAP1 in NFE2L2 mutant samples is upregulated whereas downstream targets such as NQO1, GCLC, GCLM, and others appear highly expressed indicating that the mutant NFE2L2 is insensitive to repression by KEAP1 and remains highly active, thus the GOF call.

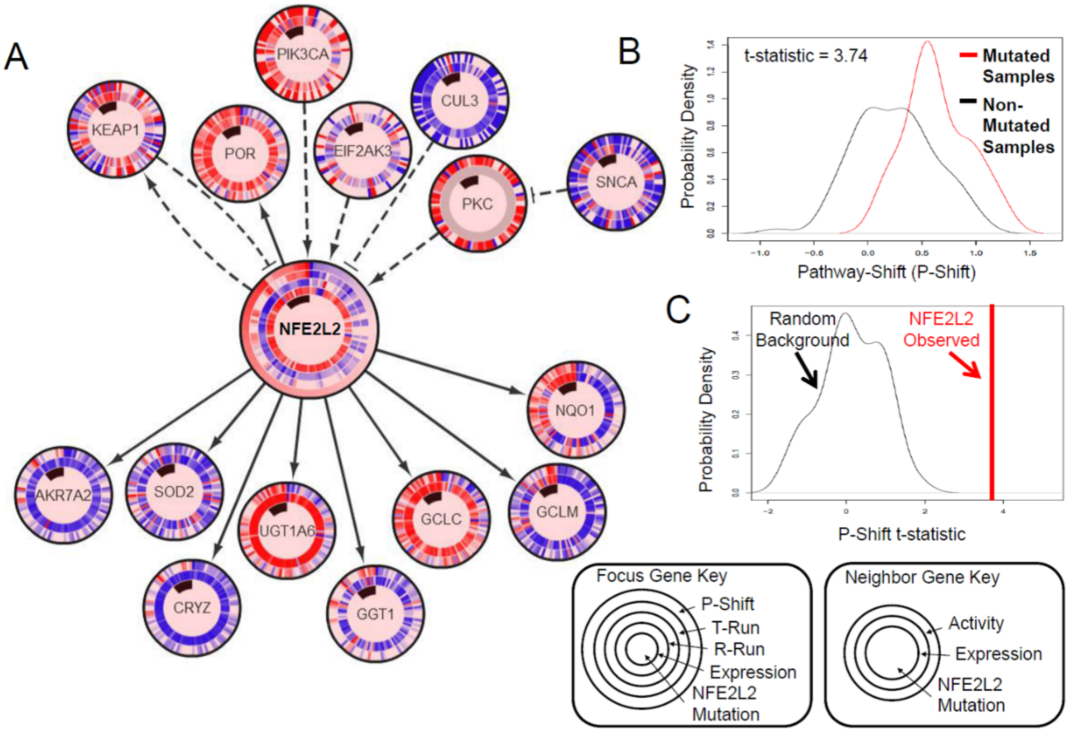


Figure 12 ***PARADIGM-SHIFT*** **result for NFE2L2 in LUSC.**

This result is illustrated in the CircleMap display in Figure 12A. The advantage of a CircleMap display is that multiple data types can be represented for a given gene. For example, NFE2L2 has five rings the inner ring indicates the samples with NFE2L2 mutation in black and samples without black as NFE2L2 wild-type. The corresponding data going outward represent the data for these two groups, expression, inferred activity from the R-run, inferred activity from the T-run, and *PS* score. The samples are sorted first by NFE2L2 mutation status then by *PS* score. The other rings have the same inner ring, but have the corresponding expression and *PARADIGM* IPL for the gene plotted. The trends to notice for a *PARADIGM-SHIFT* figure such as this is the enrichment of red, positive, *PS* scores tracking with NFE2L2 mutation versus in the wild-type indicating a GOF. This *PARADIGM-SHIFT* result is supported by high expression of NFE2L2's downstream transcriptional targets tracking with NFE2L2 mutation and high activity of NFE2L2's repressors and low activity of NFE2L2's activators upstream.

The *PARADIGM-SHIFT t*-statistic can be derived, by comparing the distribution of *PS* scores for mutant and wild-type NFE2L2 samples, Figure 12B. Random permutation analysis confirmed that the positive *PS* scores seen for mutated cases relative to the non-mutated cases were significant, Figure 12C. Thus, the method was able to predict a positive increase in activity of this gene relative to its regulatory inputs consistent with the known oncogenic influence of these mutations.

3.4 RHOA in TCGA STAD

As RHOA appears to be a highly recurrent novel mutation identified in the genome-stable (GS) molecular subgroup of TCGA stomach adenocarcinomas (STAD), to investigate the pathway evidence for loss-of-function, gain-of-function, or neutrality of specific mutations in this gene across the cohort we employed *PARADIGM-SHIFT*. Along with mutations in RHOA, we also analyzed fusion events in ARHGAP26 and ARHGAP6 as these events were found to be mutually exclusive to mutations in RHOA and also enriched in the GS cluster. The location of these mutations occur in a hotspot suggesting gain-of-function, however, the mutations were not analogous to oncogenic mutation in other RAS-family GTPases. Thus, pathway signatures identified by *PARADIGM-SHIFT* shared with the RHOA mutations could provide orthogonal evidence about the mechanism of action of these fusion events.

There were 50 samples within the GS group with available copy number and expression data to run *PARADIGM-SHIFT* analysis, with 6 RHOA mutations and 8 CLDN18-ARHGAP fusions in this set. *PARADIGM* parameters were trained on the complete cohort of samples with available copy number and expression data with a total of 258 samples. The RhoA-ROCK signaling pathway was constructed from MetaCore^{TM} and RHOA mutation neighborhoods were selected in a supervised fashion by selecting features based on *t*-statistic. The accuracy of the model is then assessed by using the absolute *PS* score as a classifier to predict the presence of an alteration (RHOA mutation or ARHGAP fusion). The model was able to predict alteration status with an average AUC of 0.62 across 5-fold cross-validation suggesting that the *PARADIGM-SHIFT* model was able to distinguish samples altered in this pathway.

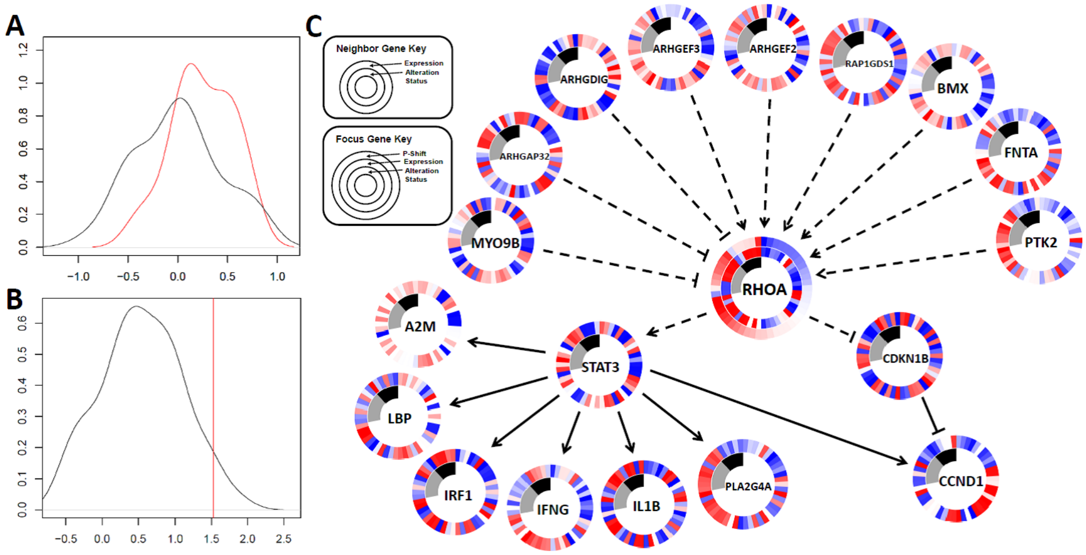


Figure 13 *PARADIGM-SHIFT* result for RHOA in STAD.

When the distribution of *PS* scores for samples with alterations in either RHOA or ARHGAP are compared to samples without either of these alterations, an enrichment of positive *PS* scores was identified indicating gain-of-function (GOF) on average through the RhoA signaling pathway, Figure 13A. Mutations are shown as black and fusions are shown in grey in the inner ring. The significance of this aggregated GOF score was determined by running a background model in which the selected network topology is fixed, but the data is permuted, thus assigning random genes to the surrounding network neighborhood of the RHOA protein. Under this background model, the GOF aggregated score was found to have a *p*-value of 0.047, Figure 13B. Altogether, these findings suggests that the signaling consequences of RHOA mutations or ARHGAP fusions lead to GOF based on the discrepancy of up- vs. down- stream activity signals.

PARADIGM-SHIFT was run on the complete cohort to determine the functional impact of alterations on the network and its network was viewed with a CircleMap display, Figure 13C [Wong et al, Nucleic Acids Research 2013]. As expected from prior knowledge, RHOA activation is mediated through the transcription factor STAT3. The pattern of expression for downstream targets IRF1 and IL1B mirrors the profile of *PS* scores concordant with RhoA pathway activation in the samples with alterations. Interestingly, downstream targets IFNG and PLA2G4A appear to be active in the case of either RHOA mutation or ARHGAP fusion. This suggests that different alterations in the RhoA pathway may not be equivalent leading to slightly different phenotypes. Additionally, the presence of samples with high *PS* scores in the non-altered set also suggests that there may be additional events within the GS subgroup not accounted for that lead to RhoA signaling activation.

3.5 TCGA PanCancer 12

High TP53 mutation rates characterize several tumor types including those represented by the three different PanCan-12 COCA subtypes C4-basal-breast, C9-ovarian, and C2-squamous-like. *PARADIGM-SHIFT* was utilized to predict the functional impact of mutations for these tumors. With the high prevalence of truncating mutations, mutations leading to the addition of a stop codon or splice site mutation, it is believed that many of the mutations in TP53 can be expected to lead to LOF. In this analysis, *PARADIGM-SHIFT* was trained on the truncating samples versus wild-type with the intention of extending the inference out to the missense mutations in order to observe any differences in TP53 pathway impact across the different types of tumors with high TP53 mutation.

Surprisingly, our pathway and gene program analysis revealed a strong prediction of LOF in C9-ovarian and C4-basal-breast, but not in C2-squamous-like. *PARADIGM-SHIFT* analysis predicts loss-of-function of TP53-truncating mutations (observed in 43% of C4-BRCA/basal, 38% of C9-OV, and 30% of C2- squamous-like cases) at a significantly higher degree in the C4-BRCA/basal and C9-OV subtypes compared to the C2- squamous-like subtype. Also, the copy-number data when aligned with TP53 missense and truncating mutations reveals more loss of heterozygosity (LOH) in the C9-OV and C4-BRCA/basal than in the C2-squamous-like samples. These observations are displayed in Figure 14 with the average *PS* scores for the truncating mutations for each cluster shown. These observations suggested a mechanism within C2-squamous-like tumors that lead the surrounding pathway to remain relatively functionally intact compared to C9-ovarian and C4-basal-breast.

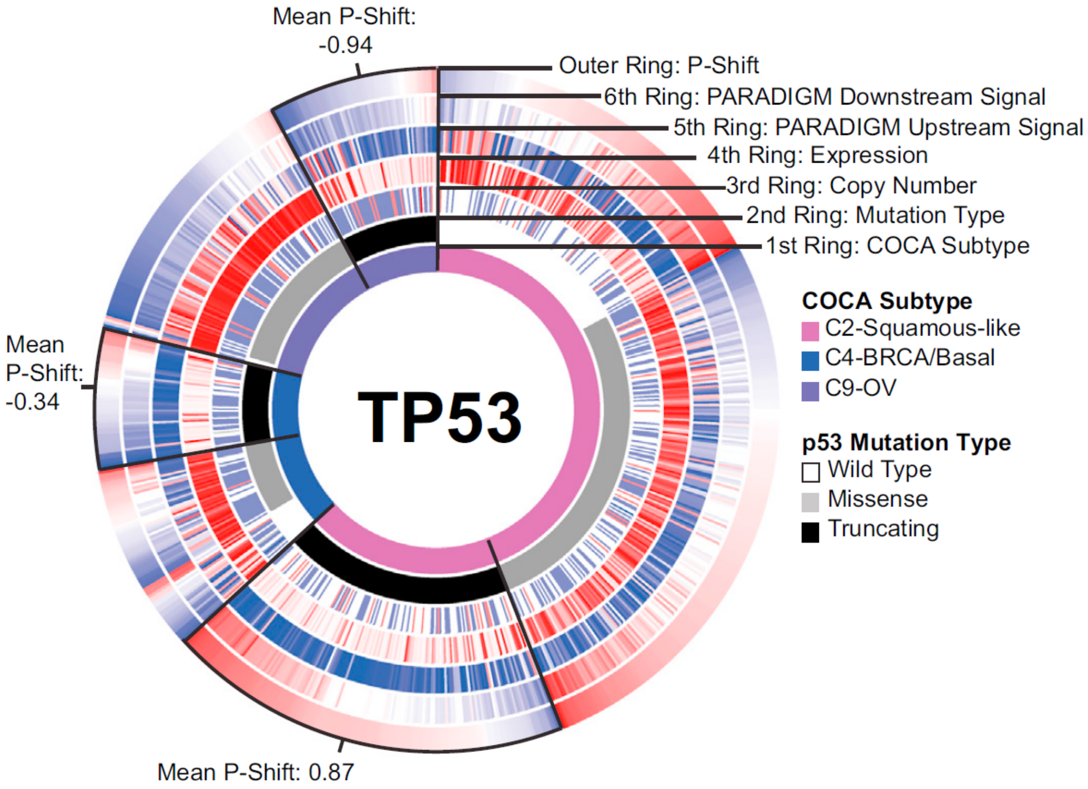


Figure 14 ***PARADIGM-SHIFT*** **result for TP53 in ovarian, basal breast, and squamous tumors in the PanCancer 12.**

Upon further investigation of the differences in the clusters identified by the *PARADIGM-SHIFT* analysis, the apparent higher TP53-pathway activity in C2-squamous-like tumors may be related to the expression of isoforms of related family members TP63 and/or TP73, which may compensate for TP53 mutation in the C2-squamous-like tumors. Notably, the transcriptional targets of TP53 shared with TP63/73 appear to be more highly expressed in the C2-squamous-like than in the C9-

ovarian or C4-basal-breast clusters. A similar finding is supported by functional experimental data in HNSC lines and tumors [Lu et al., 2011]. In HNSC, the function of TP63/73 in growth of HNSC was modulated in the presence of inflammatory factor TNF-a and cREL.

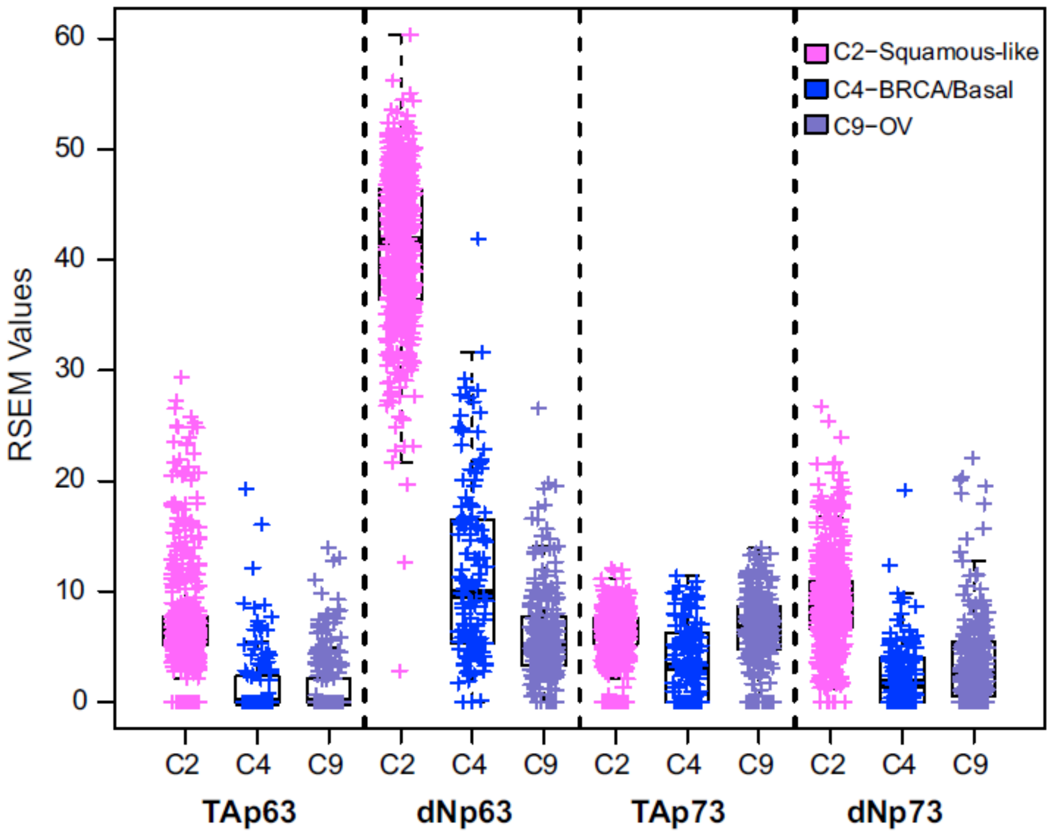


Figure 15 Isoform specific expression of TP63 and TP73 isoforms in ovarian, basal breast, and squamous tumors in the PanCancer 12.

Indeed, TP63 expression levels, in particular expression of the oncogenic dNp63 isoform, are significantly higher in the C2-squamous-like subtype than in the C4-basal-breast tumors, Figure 15. TP63 network activity or increased expression in the C9-ovarian subtype was not observed. These studies show the potential for p63/73 compensatory function for mutated or suppressed p53 in HNSC and certain breast cancer tumors, which has potential implications for targeted and standard therapy across these malignancies. These data indicate that TP53/63/73 downstream activities are of potentially broader significance among the C2-squamous-like, C9-ovarian and C4-basal-breast subtypes, with similarly high TP53 mutation rates.

3.6 NOTCH1 in TCGA LGG

NOTCH1, a highly recurrent mutation identified in the IDHmut-codel subgroup, was investigated for pathway evidence for loss-of-function, gain-of-function, or neutrality of specific mutations in this gene across the lower grade glioma (LGG) cohort. NOTCH1 has been known to harbor both activating (as in liquid malignancies) and inactivating types of mutations (as in squamous cell cancers), so *PARADIGM-SHIFT* was employed for this analysis. Thus, pathway signatures identified by *PARADIGM-SHIFT* shared with the NOTCH1 mutations could provide clues about the mechanism of action of these genomic events. The initial hypothesis for NOTCH1 in LGG was that it would likely be GOF, however LOF was predicted by *PARADIGM-SHIFT*.

There were 246 samples within the IDH mutant and wild-type subtypes and 1 sample that was not assigned a subtype (NOTCH1 fusion) with available genomic data to run *PARADIGM-SHIFT* analysis, with 29 NOTCH1 mutations, one homozygous deletion, one genomic rearrangement, and two fusions in this set. *PARADIGM* parameters were trained on the complete cohort of samples with available copy number and expression data with a total of 272 samples. The

NOTCH1 mutation neighborhoods were selected in a supervised fashion by selecting features based on *t*-statistic. When the distribution of *PS* scores for samples with alterations in NOTCH1 are compared to samples without either of these alterations, an enrichment of negative P-Shift scores was identified indicating LOF on average through the NOTCH1 signaling pathway. The significance of this aggregated LOF score was determined by running a background model in which the selected network topology is fixed, but the data is permuted, thus assigning random genes to the surrounding network neighborhood of the NOTCH1 protein. Under this background model, the LOF aggregated score was found to have a *p*-value of 0.02. Altogether, these findings suggest that the signaling consequences of NOTCH1 genomic events in LGG lead to LOF based on the discrepancy of up- vs. down- stream activity signals. *PARADIGM-SHIFT* was run on the complete cohort to determine the functional impact of alterations on the network and its network was viewed with a CircleMap display, Figure 16. The pattern of expression for many of the downstream targets of NOTCH1 mirrors the profile of *PS* score concordant with NOTCH1 pathway deactivation in the samples with alterations. Low *PS* scores are also observed in many of the other samples within the IDHmut-codel subtype, which suggests there may be additional mechanisms of NOTCH signaling pathway deactivation not considered in this analysis.

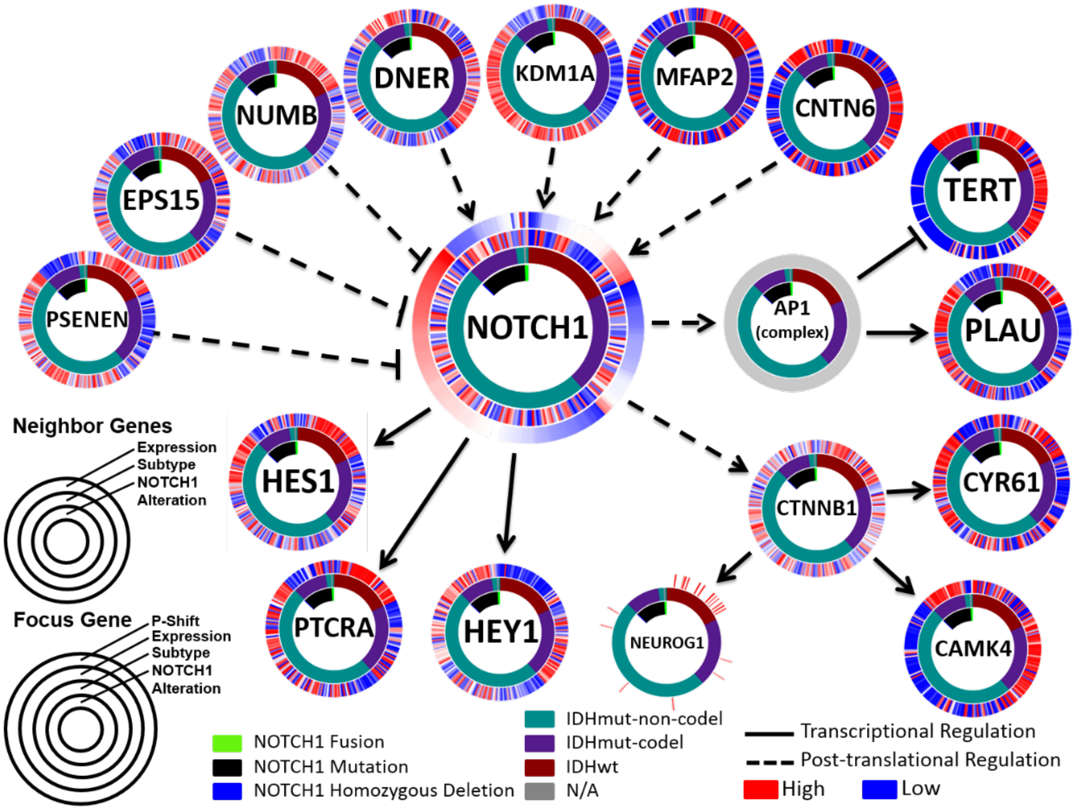


Figure 16 ***PARADIGM-SHIFT*** **result for NOTCH1 in LGG.**

Follow-up analysis of the mutational profile of alterations in NOTCH1 in LGG showed a spread of alterations in the EGF-like domain repeats as well as the intracellular domain, Figure 17A. This finding is most similar to the pattern of mutations in NOTCH1 observed in head and neck squamous carcinoma (HNSC), Figure 17B, as opposed to the pattern observed in acute lymphoblastic leukemia (ALL), Figure 17C. This supports NOTCH1 as LOF in LGG as LOF mutations are also observed in HNSC, while GOF mutations are typically observed in ALL.

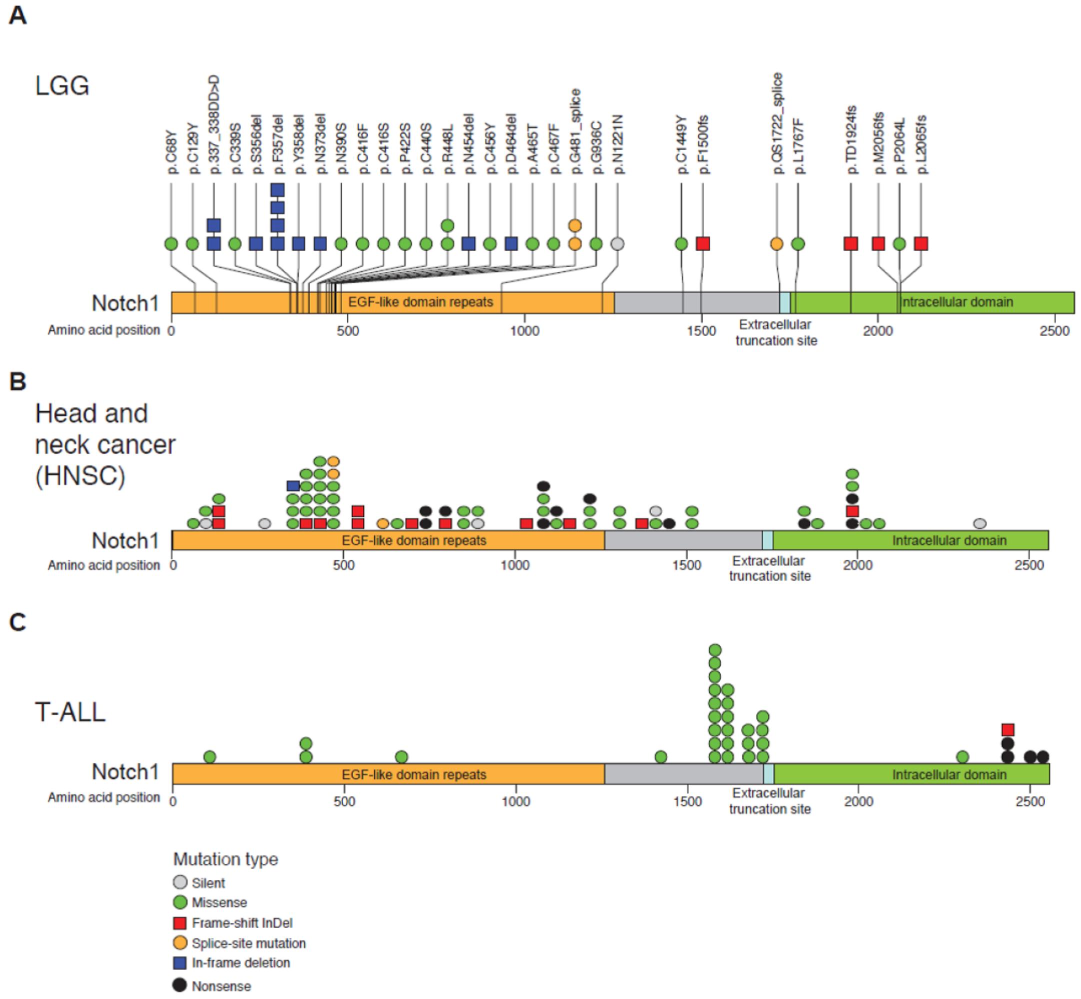


Figure 17 Mutational pattern of NOTCH1 in LGG, HNSC, and ALL.

3.7 Summary

Unlike traditional methods of assessing mutations, *PARADIGM-SHIFT* allows us to probe the genomic consequences of various genomic alterations by observing an effect on the surrounding network. This offers an orthogonal view to functional impact prediction as well as a glimpse into the mechanism by which the pathways are altered. *PARADIGM-SHIFT* has proven to be a useful tool applicable across many different tumor types assessing the impact of alterations on a specific set of alterations with network information. By assessing its pathway effect, the functional impact of mutation (GOF or LOF) was confirmed for many alterations; however, several unexpected findings were discovered as well. The lack of a predicted LOF of TP53 in squamous-like tumors across the PanCancer 12 revealed a mechanism of TP63 compensation. The presence of canonical mutations, therefore, does not necessarily mean a similar pathway affect across different types of tumors. NOTCH1 that has been documented to harbor activating and inactivating mutations was discovered to be LOF in LGG similar to the types of alterations observed in HNSC. These findings were made possible by probing the surrounding neighborhood around the affected gene and has the potential to elucidate impact in cases where the mechanism of action of mutation is not well understood.

PARADIGM-SHIFT has been shown to be a useful method, however, there are several limitations of and caveats to the analysis as well. While extreme (absolute) *PS* scores show a good overall correlation with *MutSig*, several of the genes seem to have predictions on average in the opposite direction than expected (e.g. NF1 in GBM). Complex regulatory logic surrounding the gene may show a discrepancy but the direction of the discrepancy may not always be clear. It will take further investigation into these cases to determine if a reliable direction can be inferred from the sign of the *PS* score. It may be the case that certain parts of the pathways are driven by additional alterations not accounted for in the models.

Additionally, the method can only be applied to genes with sufficient representation in the curated set of pathway interactions. While current pathway databases have a biased coverage of cancer-related genes, many of the genes with low-frequency mutations are still among those with little pathway information. It is critical to expand pathways beyond the curated set to encompass such orphan genes into the analysis of mutation consequences. We expect that methods that can computationally predict reliable casual gene-gene interactions from functional genomics datasets, such as from ChIP-Seq data with information about novel transcription-factor to target relationships, will significantly improve the breadth of genes to which pathway-based mutation impact approaches can be applied.

4 PARADIGM-SHIFT ‘Driver Modules’

One unique advantage of *PARADIGM-SHIFT* analysis of mutational impact is that, unlike most traditional methods, pathway impact can also be predicted for samples without mutations. Through several analyses in many of the TCGA working groups, we have noticed that in many cases a predicted functional impact can occur in a large proportion of samples that are wild-type for the focus gene (FG). In addition, in several cases known alterations affecting the same pathway have been shown to highly correlate with prediction of this functional impact. In this chapter, I will present work from a variety of TCGA projects illustrating the ability to predict the net functional impact of multiple events on a single pathway. While this approach is effective for confirming the net effect of known alterations on a pathway, I devise a method to infer these events based on a significant correlation to the *PS* scores in the wild-type samples.

4.1 HPV infection and TP53 in TCGA HNSC

HPV infection is a feature common to many head and neck squamous cell carcinomas. In addition, HPV infection and TP53 mutation, which are each known to lead to deactivation of the p53 signaling pathway by different mechanisms and in many cancers, appear to be highly mutually exclusive. To investigate the pathway evidence for the effect of HPV infection and TP53 mutation across the cohort, we applied *PARADIGM-SHIFT*. *PARADIGM-SHIFT* assigns a pathway impact score based on the discrepancy in upstream and downstream signal stemming from events such as mutation, but is not limited to predicting impact on samples with a mutation.

Thus, pathway signatures identified by *PARADIGM-SHIFT* shared with the TP53 mutations could provide clues about the mechanism of action of HPV infection.

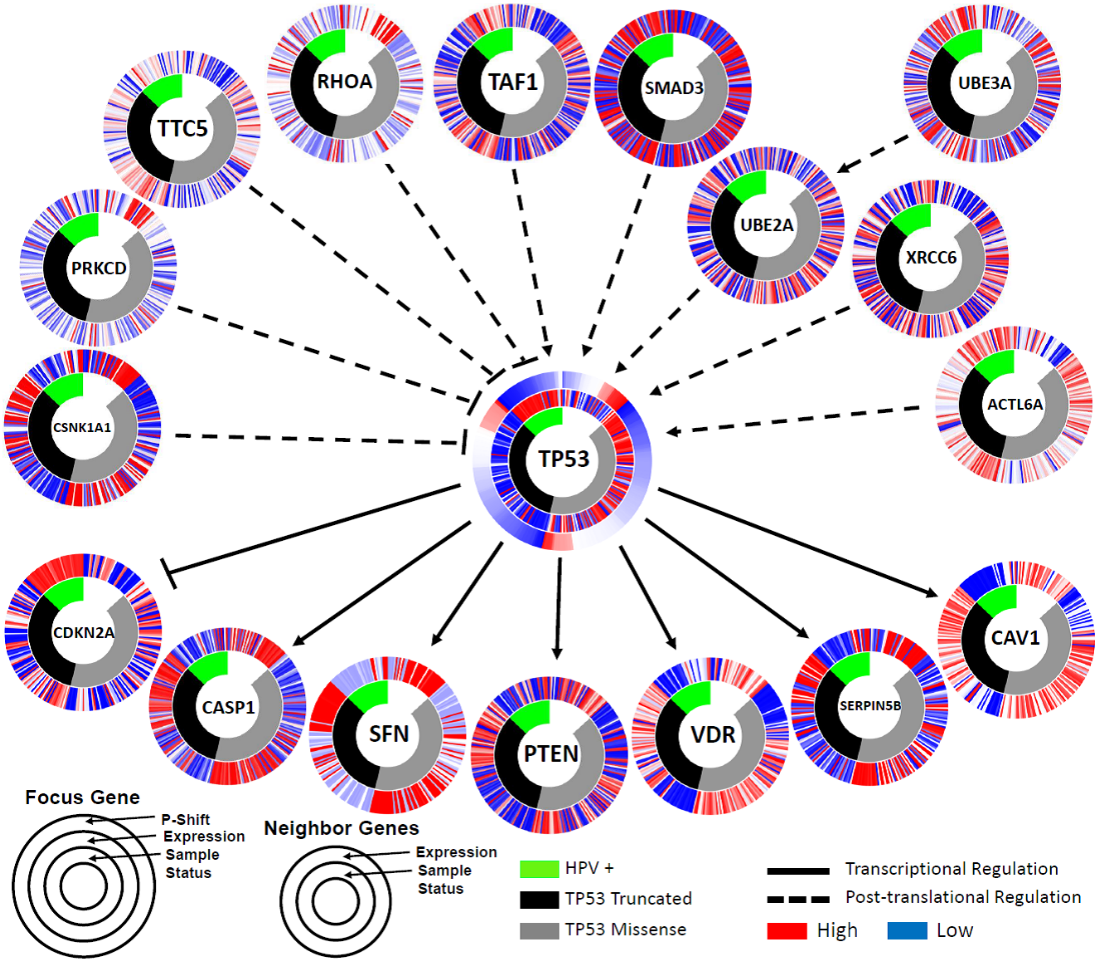


Figure 18 *PARADIGM-SHIFT* result for TP53 and HPV+ in HNSC.

There were 279 samples within the HNSC cohort having available copy number and expression data to run *PARADIGM-SHIFT* analysis, with 206 TP53 mutations and 36 HPV infected cases in this set. *PARADIGM* parameters were trained on the complete cohort of samples with available copy number and expression data with a total of 282 samples. When the distribution of *PS* scores for samples with a mutation in TP53 or HPV infection are compared to samples without either of these events, Figure 18, an enrichment of negative *PS* scores was identified indicating loss-of-function (LOF) on average through the tp53 signaling pathway. The significance of this aggregated LOF score was determined by running a background model in which the selected network topology is fixed, but the data is permuted, thus assigning random genes to the surrounding network neighborhood of the TP53 protein. Under this background model, the LOF aggregated score was found to have a p-value of < 0.0001. Altogether, these findings suggest that the signaling consequences of TP53 mutations or HPV infection lead to LOF based on the discrepancy of up- vs. down-stream activity signals. *PARADIGM-SHIFT* was run on the complete cohort to determine the functional impact of these events on the network and its network was viewed with a CircleMap display, Figure 18.

As expected, both missense and truncating mutation in TP53 appear to deactivate the p53 signaling pathway, the expression of the downstream targets of TP53 are lower in most of the mutated cases in comparison to wild-type. Interestingly, in the case of HPV positive cases the downstream targets of TP53 appear more strongly down-regulated. Additionally, regulators of TP53 activity, such as TAF1 and SMAD3, appear to be trying to activate TP53 and TP53 itself has high expression in HPV positive cases. This finding is consistent with HPV infection interfering with wild-type TP53 protein to prevent activation of its downstream targets, and succeeding in producing this net pathway LOF perhaps even more effectively than seen in cases without HPV infection but with TP53 mutation.

4.2 RTK Fusions in TCGA PanCancer 28

Mutations in receptor tyrosine kinase (RTK) genes, such as EGFR, FGFR, or c-Met, occur frequently in cancer. Alterations in RTK family genes were examined in 28 publically available TCGA tumor types, to be referred to as the PanCancer 28 set. Within the PanCancer 28 a majority of the tumor samples had an alteration in an RTK gene or in a gene involved in the RTK pathway. However, of the 28 tumor types in the PanCancer 28, 22 had sufficient genomic data to perform *PARADIGM* analysis, and of those 18 different tumor types harbored at least one alteration related to the RTK pathway. These tumor types are BLCA, BRCA, CESC, COAD, GBM, HNSC, KIRP, LGG, LIHC, LUAD, LUSC, OV, PAAD, PRAD, SARC, SKCM, THCA, and UCEC consisting of a total of 4700 tumor samples. Samples were classified into six groups based on their RTK alteration status: RTK fusion, RTK amplification, RTK mutation, mutation in a downstream signaling component of the RTK pathway, multiple RTK aberrations, and wild-type. *PARADIGM-SHIFT* was used to investigate the net pathway evidence for loss-of-function, gain-of-function, or neutrality of these alterations across these tumors. The RTK pathway is combined from regulators and targets of EGFR and FGFR, two of the main RTK genes altered in cancers.

When the distribution of *PS* scores for samples with an alteration in RTK pathway genes are compared to samples without these alterations, an enrichment of negative *PS* scores was identified indicating gain-of-function (GOF) on average through the RTK signaling pathway. *PARADIGM-SHIFT* was run on the complete cohort to determine the functional impact of these events on the network and its network was viewed with a CircleMap display, Figure 19. The type of RTK alteration is broken out in the center ring, while the average *PS* score for each group is represented on the outer ring with the average value annotated. Of the RTK alterations, RTK amplification as well as multiple aberrations and aberrations in a downstream signaling component appear the most shifted towards GOF, though RTK mutation and fusion also have average scores above those of the wild-type samples.

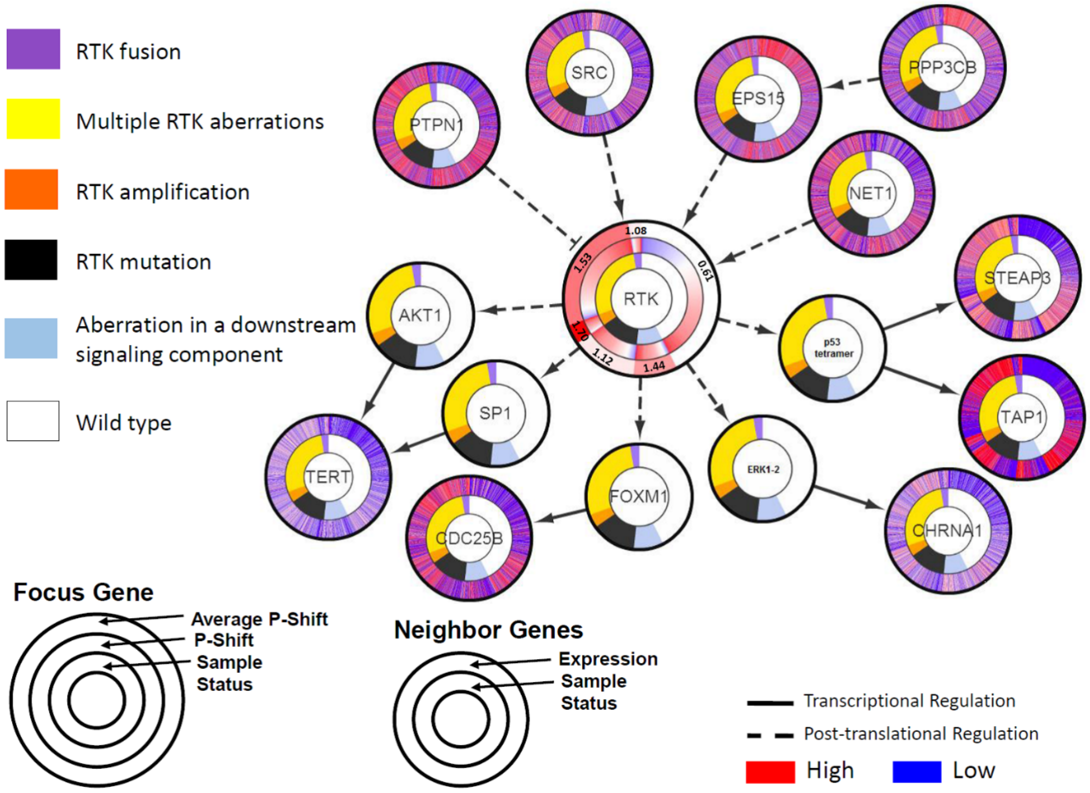


Figure 19 *PARADIGM-SHIFT* result for RTK events in the PanCancer 28.

The impact of gene fusions on RTK fusions are of particular interest to us since certain gene fusions in RTK genes have shown a strong activation in certain cohorts, such as GBM or LGG, but known fusions that disrupt the tyrosine kinase domain may be ambiguous as to whether they are functional or not. In Figure 20, RTK fusions are separated further into alterations in specific RTK families and whether or not the gene fusion preserved the entire kinase domain or interrupted it. Consistent with expectation, gene fusions that do not interrupt the kinase domain show a stronger signal for GOF based on average *PS* score compared to kinase domain disrupting counterparts. Also notably, the average *PS* score for ERBB family RTKs, such as EGFR, show very strong activation at a similar level to RTK amplifications in the previous figure. Overall these results support the conclusion that alterations in RTK pathway genes lead to activation of the RTK signaling pathways, however, certain alterations show a stronger signal of alteration such as ERBB kinase-preserving fusions and RTK amplifications.

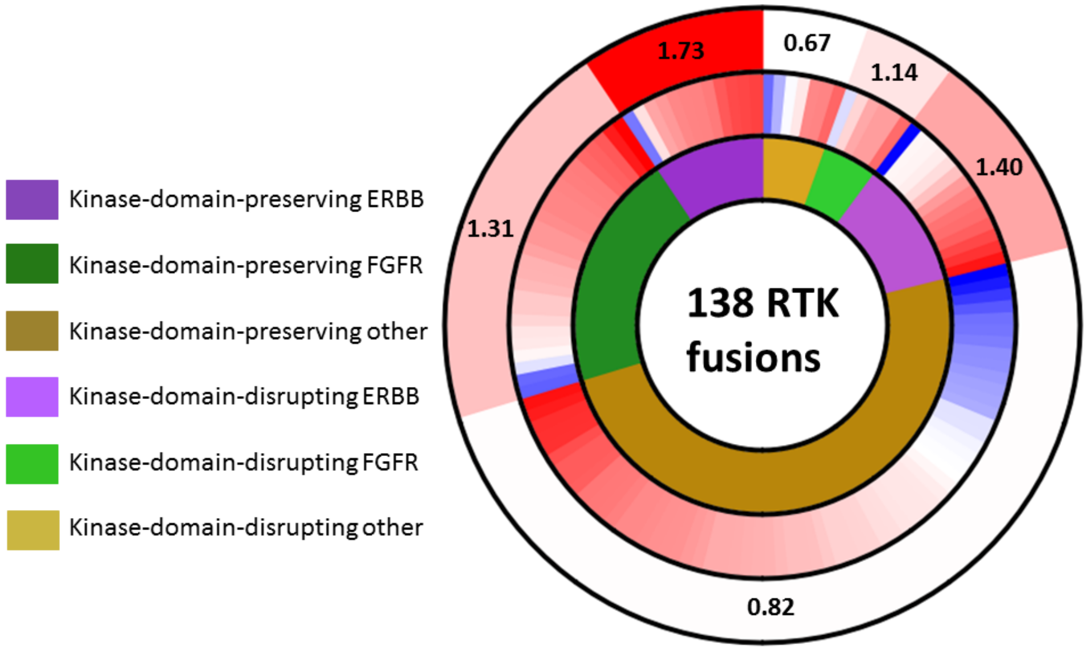


Figure 20 *PARADIGM-SHIFT* result for RTK fusion events in the PanCancer 28.

4.3 BRAF, RAS, and NF1 in TCGA Melanoma

As mutations in BRAF, RAS, and NF1 are highly recurrent and show mutual exclusivity in melanoma tumors, *PARADIGM-SHIFT* was employed to investigate the pathway evidence for loss-of-function, gain-of-function, or neutrality of these mutations across the cohort. Samples available for analysis that had both copy number and expression data included 118 with BRAF hotspot mutations, 81 with RAS hotspot mutations, 23 with any NF1 mutation, and 38 without any of these mutations (triple wild-type). *PARADIGM-SHIFT* was used to identify pathway signatures shared across these mutations that all impinge on the BRAF-MAPK signaling pathway, which could provide clues about the mechanism of action of these genomic events.

PARADIGM parameters were trained on 335 samples that had available copy number and expression data. Pathway neighborhoods for BRAF were constructed before running the algorithm to find those most informative for predicting gain- or loss-of-function based on the activities of the surround regulators and targets. Genes in the BRAF neighborhood were selected in a supervised fashion based on a *t*-statistic score that included genes that showed differential expression for BRAF mutant versus triple wild-type greater than one standard deviation above the mean. The same neighborhoods were used for the RAS mutant and NF1 mutant versus triple wild-type calculations. *PARADIGM-SHIFT* (*PS*) scores for BRAF, which reflect the discrepancy in upstream versus downstream pathway signals, were calculated as the difference in inferred activity between the two runs of *PARADIGM*.

When the distribution of *PS* scores for samples with alterations in BRAF, RAS, or NF1 are compared to samples without these alterations, an enrichment of positive *PS* scores was identified indicating gain-of-function (GOF) on average through the BRAF-MAPK signaling pathway for these alterations. The significance of this aggregated GOF score was determined by obtaining a background model in which the selected network topology is fixed, but the data is permuted, thus assigning random genes to the surrounding network neighborhood of the BRAF protein. Under this background model, the GOF aggregated score was found to have a p-value of < 0.0001 for BRAF hotspot mutation versus triple wild-type, 0.0004 for RAS hotspot mutation versus triple wild-type, and 0.005 for NF1 any mutation versus triple wild-type, Figure 21A. As expected, these findings suggest that the signaling consequences of genomic events in these mutations in SKCM lead to GOF based on the discrepancy of up- vs. down- stream activity signals.

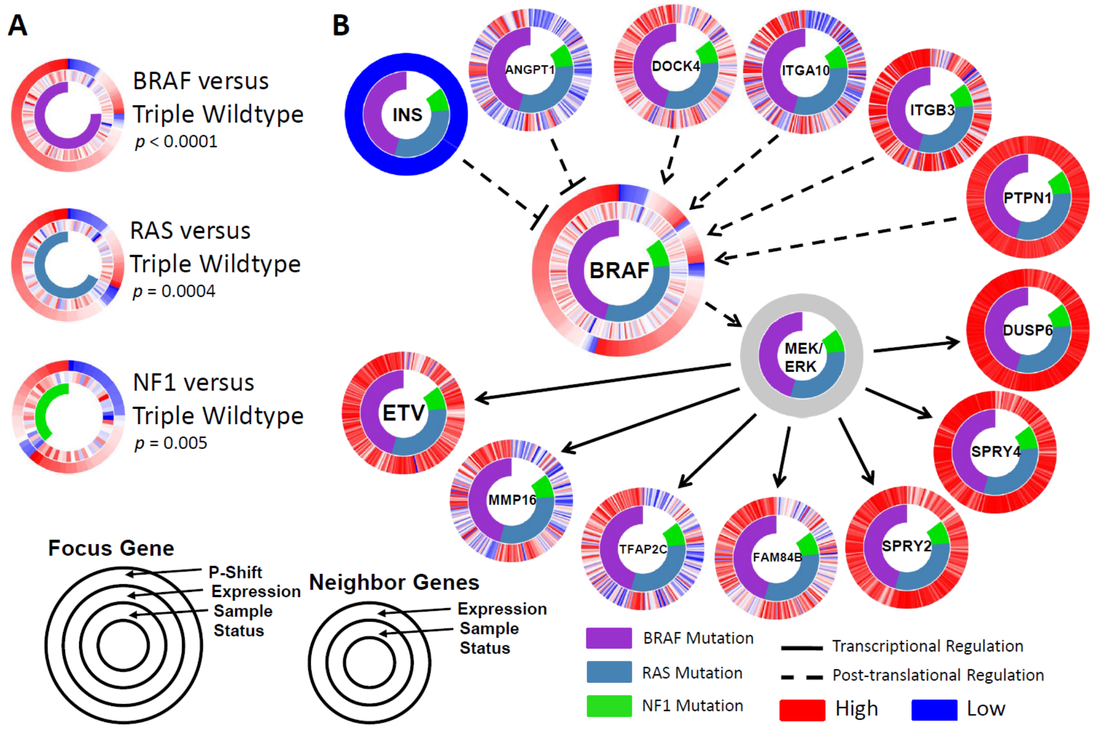


Figure 21 *PARADIGM-SHIFT* result for BRAF, RAS, and NF1 in SKCM.

PARADIGM-SHIFT was run on the complete cohort to determine the functional impact of alterations on the BRAF network and was visualized with a CircleMap display, Figure 21B. The pattern of expression for many of the downstream transcriptional targets of BRAF-MAPK signaling correlates with the profile of *PS* scores concordant with BRAF pathway activation in the samples with alterations. High *PS* scores are also observed in many of the other samples belonging to the triple wild-type group, which suggests there may be additional mechanisms of BRAF signaling pathway activation not considered in this analysis.

4.4 NFE2L2 and KEAP1 in PanCancer 12

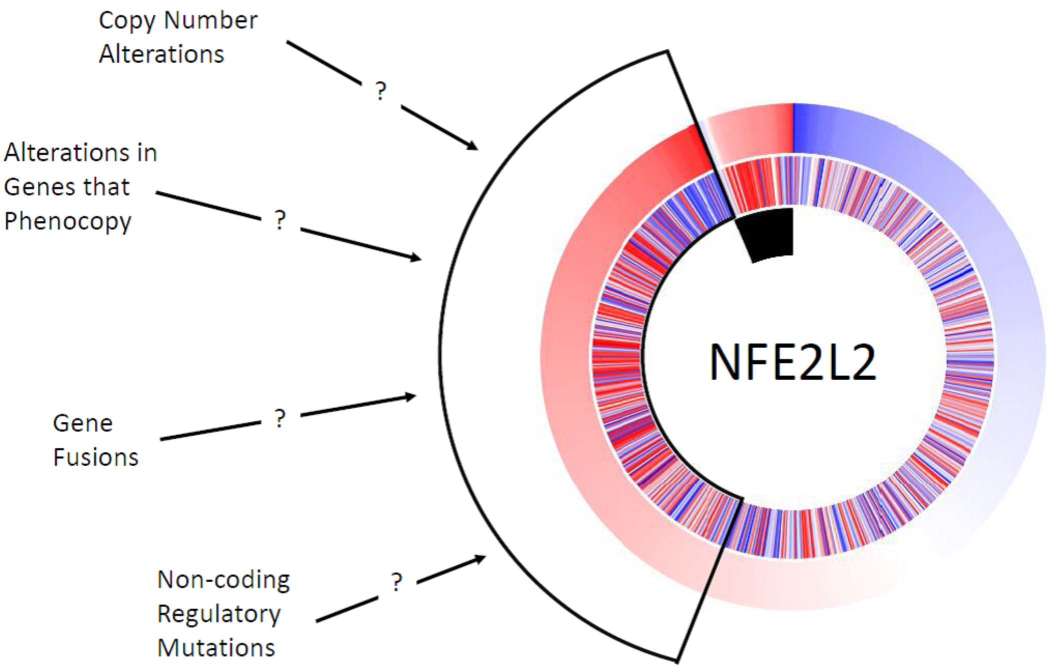
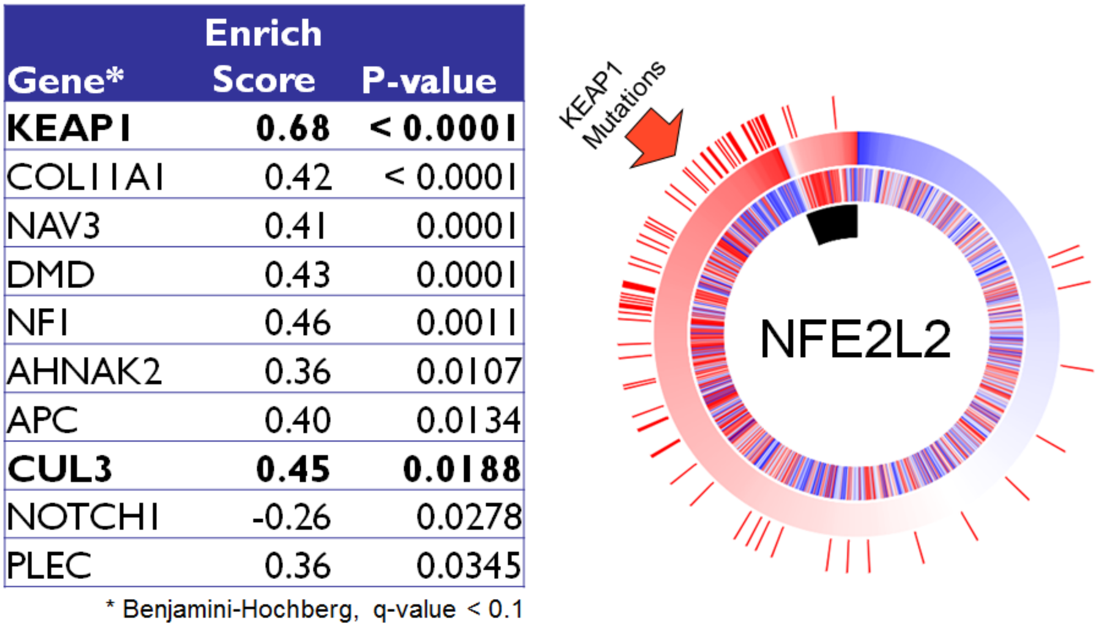


Figure 22 *PARADIGM-SHIFT* result for NFE2LE in PanCancer 12 showing samples with high PS score in many NFE2L2 wild-type cases.

While NFE2L2 mutation predicts GOF with high *PS* scores, there also appears to be many samples that do not harbor a mutation in NFE2L2 that have high *PS* scores as well, Figure 22. These samples potentially have additional alterations that could lead to activation of the same pathway through an alternate mechanism. Subsequent analyses identified KEAP1 mutations are enriched in the set of samples with higher *PS* scores, Figure 23. This analysis is performed by taking the ranked list of samples and looking for enrichment of other genomic events correlating with high predicted impact. These enrichments are calculated similar to running *GSEA*, but performed on a ranked list of samples instead of genes. The sets contain the list of samples altered for a particular gene. *GSEA* identifies which of these “event sets” is most enriched for predicted GOF or LOF. This approach outlined here allows us to identify events associated with a predicted functional impact without prior knowledge of these events. This finding of KEAP1 mutations being enriched in the NFE2L2 wild-type samples with high *PS* scores supports that KEAP1 mutation is an alternative mechanism of activating the Nrf2 signaling pathway, which is consistent with prior knowledge of Nrf2 signaling pathway activation. It is known that in wild-type cases NFE2L2 and KEAP1 normally interact to signal for NFE2L2 to be degraded, however, this finding suggests mutation of either NFE2L2 or KEAP1 disrupts this interaction therefore leading to higher activity of NFE2L2. In addition to identifying KEAP1 associations, several other genes are implicated as well which are involving in the Nrf2 signaling pathway or have evidence of interacting with NFE2L2 indirectly, such as COL11A1, NAV3, NF1, APC, and CUL3.



Gene*	Enrich Score	P-value
KEAP1	**0.68**	**< 0.0001**
COL11A1	0.42	< 0.0001
NAV3	0.41	0.0001
DMD	0.43	0.0001
NF1	0.46	0.0011
AHNAK2	0.36	0.0107
APC	0.40	0.0134
CUL3	**0.45**	**0.0188**
NOTCH1	-0.26	0.0278
PLEC	0.36	0.0345

* Benjamini-Hochberg, q-value < 0.1

Figure 23 KEAP1 mutations is most significantly associated with NFE2L2 predicted GOF in wild-type samples.

In this example, *PARADIGM-SHIFT* analysis was able to correctly identify NFE2L2 mutation as well as linking KEAP1 mutation as driving factors leading to Nrf signaling pathway activation. *PARADIGM-SHIFT* uses the difference in inferred upstream and downstream activity, which allows us to make inferences about pathway activation or de-activation regardless of whether SNV data is present. Since *PARADIGM-SHIFT* can be run on samples without alterations in the focus gene, we can look for hits to the pathway that lead to the same phenotype from copy number alterations, changes in gene expression, gene mutations and fusions, or other events. In addition, *PARADIGM-SHIFT* allows us to identify the genes that appear most affected by genomic aberrations to these pathways that may be helpful for identifying effective interventions for treating these cancers. In the next section, I outline the procedure in further detail and show an example in which the iterative discovery of associated events improves discovery of additional associated events.

4.5 The PARADIGM-SHIFT Molecular Machines Method

Predicting the functional impact of mutations from pathways is a challenging problem, first we need accurate models of the interaction network in order to utilize pathways effectively, and second, the information we have about which samples are likely altered is incomplete. There are false positive calls of mutations, mutations with no functional consequence, and mutations in genes that have a similar consequence in the same pathway. If this last group, mutations in genes that have a similar consequence in the same pathway, can be identified for many of the known *driver* pathways this has the greatest potential for extending our understanding of mutational *drivers*. While many tumors typically have mutations in several well characterized *driver* genes, there are many tumors that do not have mutations in any of these but have rarer *driver* events. Finding these "molecular machines" will be key for developing new treatments and could potentially identify patients subgroups that will respond to known therapies that target these pathways.

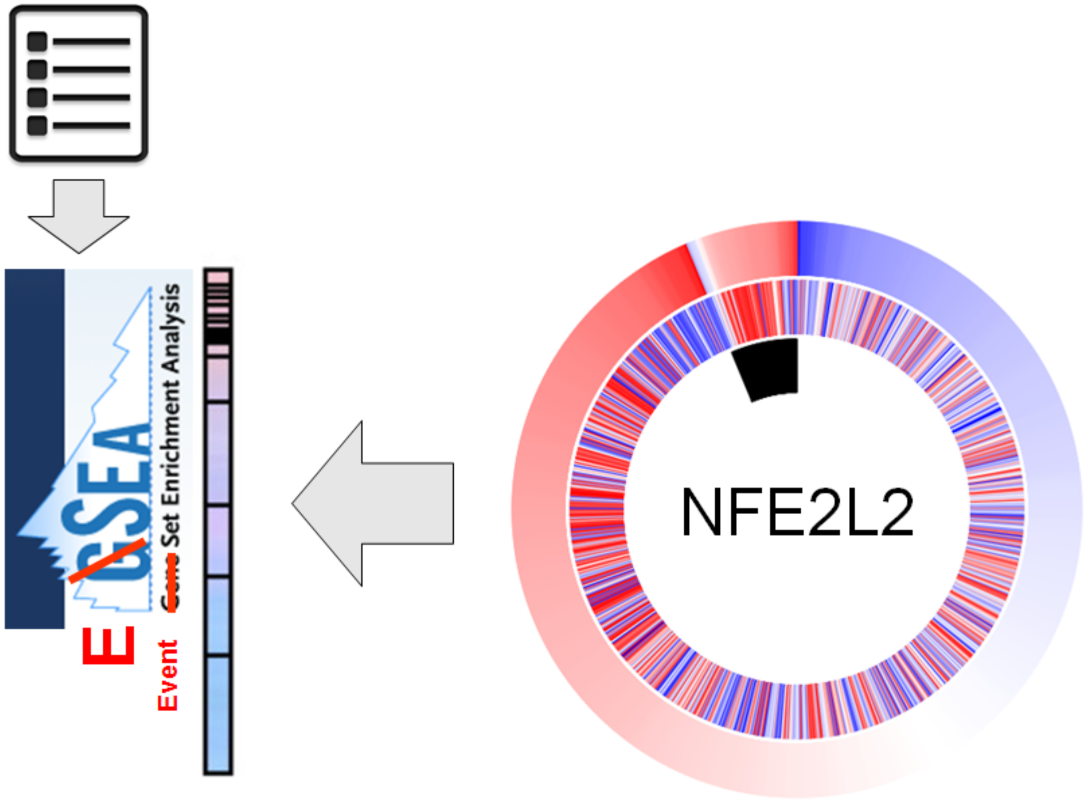


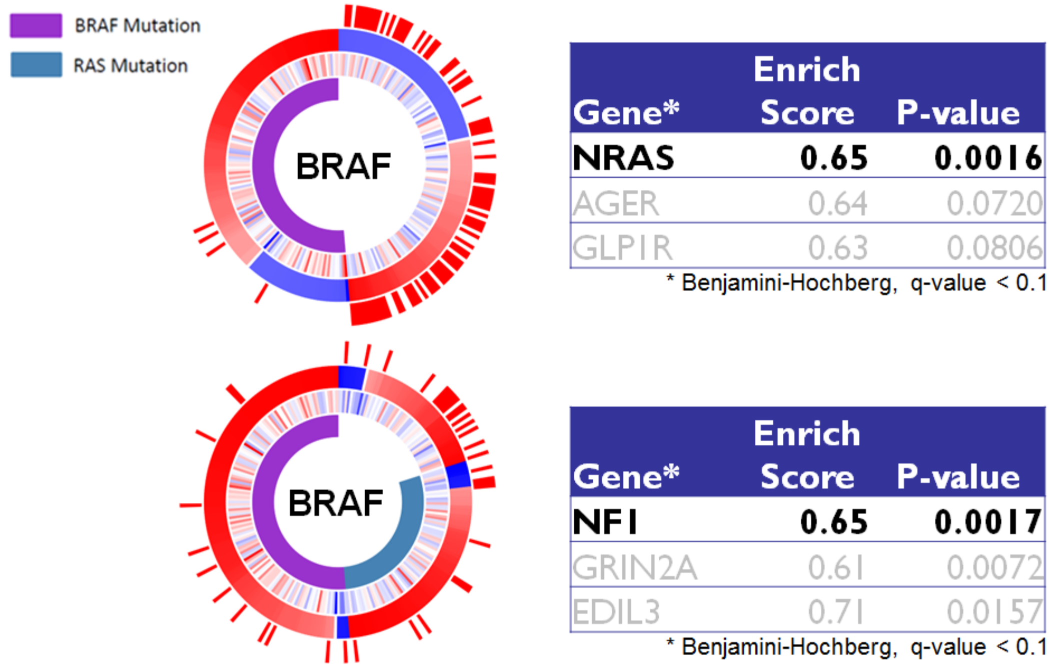
Figure 24 Procedure for discovering significantly associated events by *PARADIGM-SHIFT* score.

These events can be identified by looking for enrichment of the samples correlating to a predicted functional impact. I employ *GSEA* for this analysis, however, where *GSEA* looks for enrichment of gene sets in a ranked list of samples, I am looking for enrichment of sample sets in either tail of a ranked list of samples. *PS* score determines the sample rank and sample sets are created for whether those samples contain a particular event. *GSEA* can then be applied to discover which events are significantly associated, Figure 24. Significance is estimated by generating null permutations of each sample set and estimating the probability of obtaining an enrichment as or more significant than the true enrichment. In order to control for multiple hypotheses, the number of events tested can be limited by evidence of mutual exclusivity, pathway proximity, or predicted *drivers* by recurrence methods, such as *MutSig* or *MuSic*. This is a necessary step since there would not be sufficient statistical power to test all hypotheses. Significant events are determined by fixing an FDR rate, $\alpha = 0.1$, by the Benjamini-Hochberg procedure [Benjamini et al. 1995].

In the next section, I will show the example of BRAF GOF in SKCM again, but demonstrate that this approach for identifying “molecular machines” can be used to identify NRAS and NF1 without biasing the analysis with prior knowledge. I will also demonstrate that by pulling in the events identified into the positive set, by building up this “molecular machine” iteratively, it increases our power to detect additional events. This could be due to two major reasons. First, pulling in additional true positives, samples with a functional impacted pathway, from the negative set to the positive set allows for more accurate training of the *PARADIGM-SHIFT* network. Second, trying to identify multiple associated events reduces our power to detect since they confound each other in the enrichment; this is especially true for lower frequency events.

4.5.1 Application to BRAF in SKCM

In the previous analysis of BRAF in melanoma (SKCM), RAS and NF1 mutants were removed from the wild-type set since we had a prior expectation that these mutants were involved in BRAF signaling, however, in a true scenario in which we want to identify novel associated events, we will not have this prior knowledge. I then apply the *PARADIGM-SHIFT* “molecular machines” methodology to BRAF in SKCM, in order to test our ability to detect RAS and NF1 as associated events.



Gene*	Enrich Score	P-value
NRAS	**0.65**	**0.0016**
AGER	0.64	0.0720
GLP1R	0.63	0.0806

* Benjamini-Hochberg, q-value < 0.1

Gene*	Enrich Score	P-value
NF1	**0.65**	**0.0017**
GRIN2A	0.61	0.0072
EDIL3	0.71	0.0157

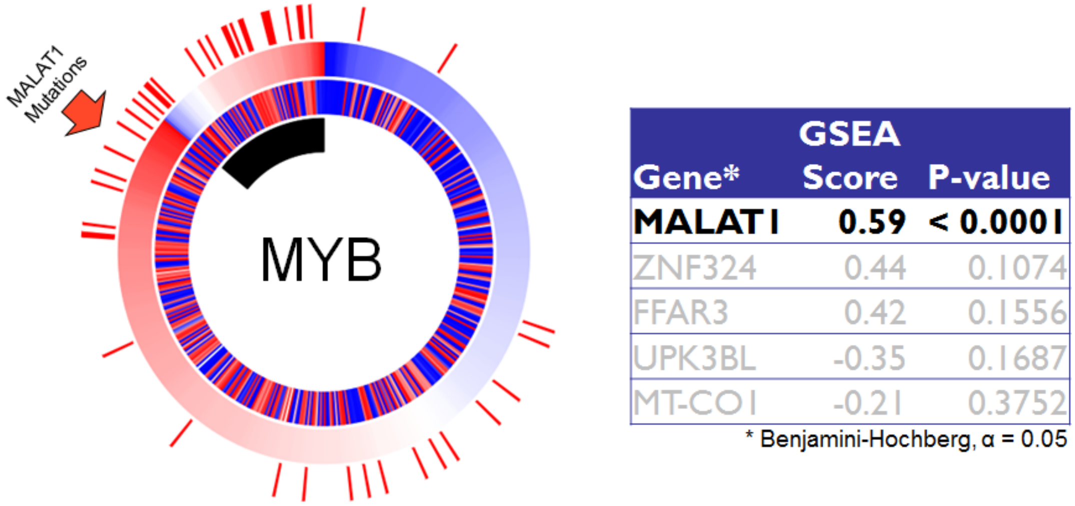
* Benjamini-Hochberg, q-value < 0.1

Figure 25 *PARADIGM-SHIFT* identifies BRAF/NRAS/NF1 molecular machine.

Since GISTIC copy number calls are available for SKCM, mutation is classified as the presence of a non-silent SNV or an absolute CNV greater than or equal to two. *PARADIGM-SHIFT* was run for BRAF mutation versus wild-type, identifying a significant GOF for BRAF signaling and identifying NRAS as significantly associated. Combining NRAS mutations into the positive set and retraining then allows us to identify NF1, Figure 25. As seen in the second PARADIGM-SHIFT run including BRAF and NRAS, the distribution of *PS* scores changes significantly. Additionally, the new model accounts for more of the BRAF mutant samples being predicted as GOF, while many wild-type samples also have high *PS* scores. This suggests that the majority of samples in SKCM have activated BRAF signaling, however, the degree of activation may differ across different mechanisms of activation.

4.5.2 MYB and MALAT1 in the Larsson 505

Along with identifying infrequent drivers, *PARADIGM-SHIFT* can be extended to events that cannot be assessed by traditional approaches, such as the impact of HPV in HNSC as previously discussed. This also includes the possibility of discovering the functional impact of non-coding mutations. Not much is known about the function of non-coding mutations and the pathways they are involved in, however, if they can be associated with a functional impact on a known *driver* pathway, we may gain a greater understanding of these mutations. In a recent effort the International Cancer Genome Consortium (ICGC) aims to characterize a large number of tumors for mutations in long non-coding RNA (lncRNA) mutations in addition to coding mutations. In a pilot study of 505 samples mutations in MALAT1, a long non-coding RNA, is found to be associated with MYB pathway GOF, Figure 26. This connection is consistent with literature demonstrating MALAT1 up-regulation leads to down-regulation of MYB [Wang et al. 2014]. This *PARADIGM-SHIFT* analysis suggests the mutations in MALAT1 are LOF due to the loss of regulation with MYB. Additionally, MALAT1 was found to be significantly mutated across multiple tumors in the original PanCancer 12 effort.



Gene*	GSEA Score	P-value
MALAT1	**0.59**	**< 0.0001**
ZNF324	0.44	0.1074
FFAR3	0.42	0.1556
UPK3BL	-0.35	0.1687
MT-CO1	-0.21	0.3752

* Benjamini-Hochberg, α = 0.05

Figure 26 *PARADIGM-SHIFT* identifies MALAT1 association with MYB GOF.

4.6 Summary

In this chapter, I have demonstrated that *PARADIGM-SHIFT* is capable of detecting net functional impacts of mutation across multiple individual events. These form what I term a “molecular machine,” a group of different events that lead to a similar pathway impact. These events are not limited to mutations in genes, but can include clinical information such as HPV infection causing TP53 LOF. By employing a sample-wise enrichment analysis with *GSEA*, I can discover events significantly associated with predicted functional impacts in samples wild-type for a given focus gene. Detection of these events can be confounded by the possibility of multiple events in the “molecular machine” along with wild-type samples not truly being wild-type for other mutations, leading to incomplete knowledge input for the training of these *PARADIGM-SHIFT* models. As a result, I show that an iterative approach may be helpful in the discovery of additional events, such as NRAS, and NF1 mutations that are associated with BRAF mutations in SKCM. Lastly, with the characterization of non-coding mutations in ICGC, a huge opportunity to relate non-coding mutations to pathways with *PARADIGM-SHIFT* is now available across a large set of tumors across many tumor types.

5 Galaxy Integration

Computational science is a rapidly evolving field with many exciting developments, however, with the growing number of computational methods and publically available datasets reproducibility is a major issue. Data types do not always have a standardized format and even issues such as how missing data is declared can cause problems for running various computational methods. In a recent article about reproducible computational research, several rules were outlined for how reproducibility can be achieved [Sandve et al. 2013].

To address these concerns I have made my code publically available and have developed them to work within the Galaxy framework, with the goal of making my computational tools more accessible as well as reproducible. In this chapter, I will overview the tools I have contributed to Galaxy and discuss the unique challenges that we continue to face using high dimensional data and the rapidly evolving environment of computational research [Boekel et al. 2015].

5.1 Addressing the Need for Accessible and Reproducible Computational Tools

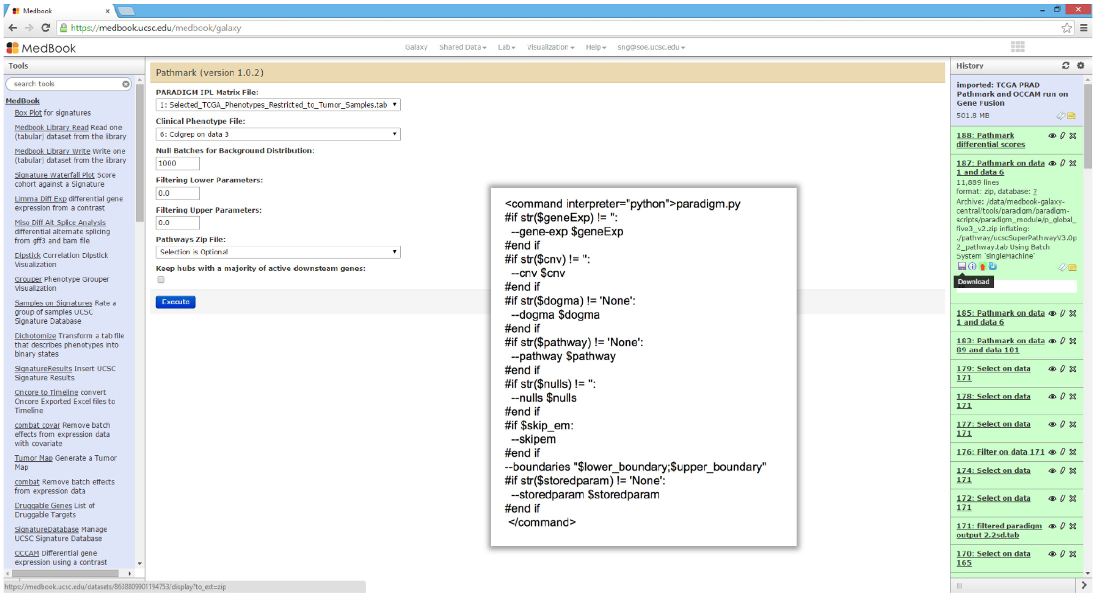
Between 2000 and 2010 there were 742 English research papers retracted from the PubMed database [Steen 2010]. Most of these retractions were due to scientific mistakes, but there has been a surprisingly high incidence rate of fraud as well. Reproducibility and accessibility of methods and data is key to ensure the validity of scientific discovery in this era of big data. In order for this to happen several rules need to be followed to ensure analytical reproducibility. Failure of replicability can occur when analytical steps between data and results are not properly recorded, such as specification of model parameters being used or data manipulation steps. Research code changes quickly and if version information is lost then older results become irreproducible. Accessibility of methods is also a challenge since setting up tools written by others is no easy task; inputs and settings that are clear to the developer can be unclear to the average user. In an effort to address these challenges, I have incorporated my tools into the Galaxy framework, a web-based platform for transparent computational biomedical research.

5.2 UCSC Pathway Analysis Toolshed

In addition to a whole suite of standard tools that come alongside a Galaxy instance, support for incorporating custom tools exists as well. Many of the tools used by the Stuart lab have been incorporated into Galaxy, from data manipulation to computational tools with report outputs. In this section, I discuss the advantages of incorporating these directly into Galaxy and describe my contributions to this UCSC Pathway Analysis Toolshed.

For a developer, it is fairly simple to take a tool that runs on the command line to one that is runnable through a Galaxy interface. With a few extra considerations about which files are output from the tools and tool dependencies, Galaxy tool development is relatively straightforward [Blankenberg et al. 2014]. Briefly, the developer describes the inputs and parameters visible to the users and indicates how those arguments are passed to the tools on the command-line running behind a Galaxy server through an XML formatted description, Figure 27. This allows users other than the developer to run tools with much more ease than on the command-line.

The inputs and parameters are made much clearer and the developer can restrict file types acceptable for inputs to reduce improper use. Additionally, with the Galaxy interface the history of analyses as well as version control of methods is available, so that any result made in Galaxy is more readily reproducible than "naked" code.


Medbook
https://medbook.ucsc.edu/medbook/galaxy
MedBook
Galaxy
Shared Data
Lab
Visualization
Help
sng@soe.ucsc.edu
Tools
search tools
MedBook
Box Plot for signatures
Medbook Library Read Read one (tabular) dataset from the library
Medbook Library Write Write one (tabular) dataset from the library
Signature Waterfall Plot Score cohort against a Signature
Limma Diff Exp differential gene expression from a contrast
Miso Diff Alt Splice Analysis differential alternate splicing from gff3 and bam file
Dipstick Correlation Dipstick Visualization
Grouper Phenotype Grouper Visualization
Samples on Signatures Rate a group of samples UCSC Signature Database
Dichotomize Transform a tab file that describes phenotypes into binary states
SignatureResults Insert UCSC Signature Results
Oncore to Timeline convert Oncore Exported Excel files to Timeline
combat covar Remove batch effects from expression data with covariate
Tumor Map Generate a Tumor Map
combat Remove batch effects from expression data
Druggable Genes List of Druggable Targets
SignatureDatabase Manage UCSC Signature Database
OCCAM Differential gene expression using a contrast
Execute
History
Download
https://medbook.ucsc.edu/datasets/8638809901194753/display?to_ext=zip

Pathmark (version 1.0.2)

PARADIGM IPL Matrix File: 1: Selected_TCGA_Phenotypes_Restricted_to_Tumor_Samples.tab

Clinical Phenotype File: 6: Colgrep on data 3

Null Batches for Background Distribution: 1000

Filtering Lower Parameters: 0.0

Filtering Upper Parameters: 0.0

Pathways Zip File: Selection is Optional

Keep hubs with a majority of active downsteam genes: ☐

```
<command interpreter="python">paradigm.py
#if str($geneExp) != ":
  --gene-exp $geneExp
#end if
#if str($cnv) != ":
  --cnv $cnv
#end if
#if str($dogma) != 'None':
  --dogma $dogma
#end if
#if str($pathway) != 'None':
  --pathway $pathway
#end if
#if str($nulls) != ":
  --nulls $nulls
#end if
#if $skip_em:
  --skipem
#end if
--boundaries "$lower_boundary;$upper_boundary"
#if str($storedparam) != 'None':
  --storedparam $storedparam
#end if
 </command>
```

imported: TCGA PRAD Pathmark and OCCAM run on Gene Fusion
501.8 MB

188: Pathmark differential scores

187: Pathmark on data 1 and data 6
11,889 lines
format: zip, database: ?
Archive: /data/medbook-galaxy-central/tools/paradigm/paradigm-scripts/paradigm_module/p_global_five3_v2.zip inflating: ./pathway/ucscSuperPathwayV3.0p2_pathway.tab Using Batch System 'singleMachine'

185: Pathmark on data 1 and data 6

183: Pathmark on data 89 and data 101

179: Select on data 171

178: Select on data 171

177: Select on data 171

176: Filter on data 171

174: Select on data 171

172: Select on data 171

171: filtered paradigm output 2.2sd.tab

170: Select on data 165

Figure 27 *PATHMARK* tool in Galaxy with a record of the analysis on the History bar.

I have incorporated all the tools discussed in this Thesis to run behind a Galaxy server, including visualization tools such as scripts for the generation of CircleMaps and Cytoscape networks.

5.2.1 PARADIGM

PARADIGM analysis lays the foundation for many of the analyses performed in our lab. It is a versatile tool that integrates knowledge from a variety of different data sources and pathways in order to make sense of which pathway features appear activated. The *PARADIGM* binary can take in data matrices as input, however, due to memory issues and speed it is not practical to run *PARADIGM* on a single process. In order to handle extremely large *PARADIGM* analyses a set of scripts are used as a wrapper for running *PARADIGM* in a parallel compute setting. I contributed to this code base by linking the code in with Galaxy, reducing the number of command-line calls and also making the scripts more efficient, so that we had the capacity to run on larger datasets. Example data are available examining inferred activities on a small pathway across a few cancer tumors.

5.2.2 PATHMARK

PATHMARK was developed to perform differential analysis on *PARADIGM*. In order to leverage the SuperPathway used by *PARADIGM*, *PATHMARK* identified clusters of highly differential *PARADIGM* features on the network. One of the major advantages of the method was its ability to produce results that are visually interpretable. *PATHMARK* analysis can be run directly in Galaxy producing files that can be visualized over the web with CytoscapeJS, discussed later, or through Cytoscape. Additional features not originally described in the original breast cancer cell lines paper were also developed for more general use, such as bootstrapping to determine the robustness of certain pathway regions or heat diffusion to propagate signal and promote the selection of a wider region. Example data are available for examining basal breast pathway markers.

5.2.3 PARADIGM-SHIFT

PARADIGM-SHIFT performs functional impact assessment of genomic alterations by integrating *PARADIGM* inference to identify discrepancies in pathway signal. By probing the effect on the surrounding pathway, *PARADIGM-SHIFT* provides a unique perspective to predict functional impact. Integration with *PARADIGM* and the visualization tools within Galaxy simplify running *PARADIGM-SHIFT* within the Galaxy environment. Example data are available for predicting NFE2L2 GOF across lung tumors.

5.2.4 CircleMaps and CytoscapeJS

As many of my tools are pathway-based, I have worked on various visualizations that aid in the display of these results. CircleMaps are particularly useful for network displays because by creating circular heatmaps, different samples and data types can be displayed for multiple genes on a network [Wong et al. 2013]. CytoscapeJS can then be used to display the network for *PATHMARK* or *PARADIGM-SHIFT* on web report allowing the results to interface directly through Galaxy. An example of a *PARADIGM-SHIFT* result displayed with CytoscapeJS with CircleMaps is shown below, Figure 28.

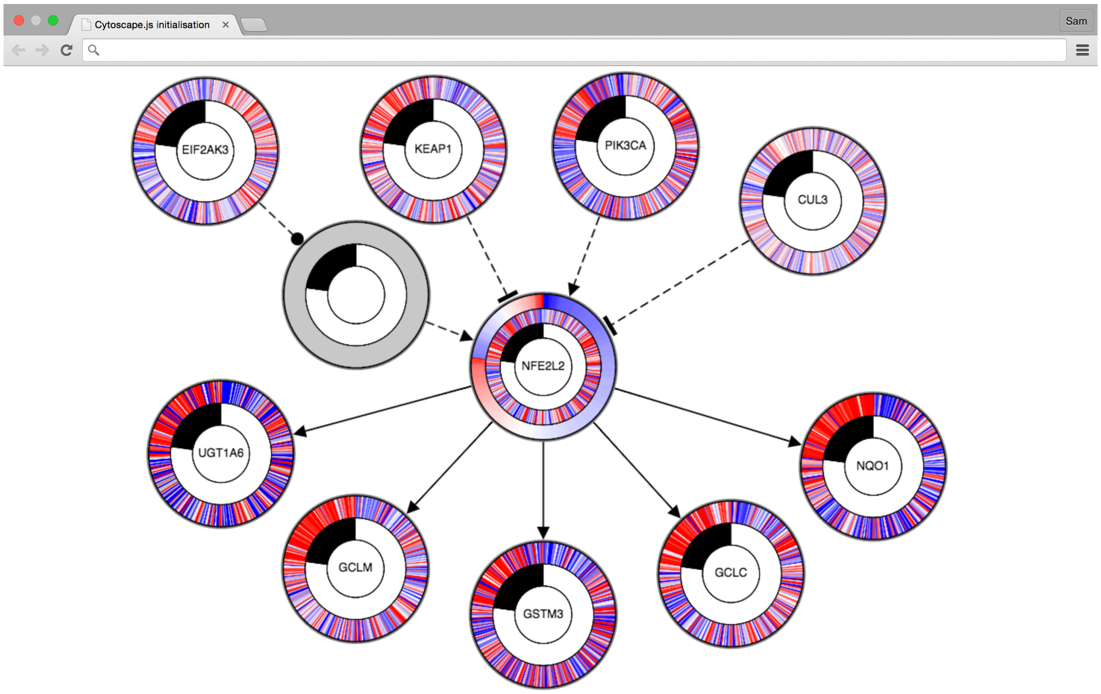


Figure 28 ***CytoscapeJS*** **view with CircleMaps displayed.**

5.2.5 Making the Connections with Galaxy Workflows

Another advantage of using Galaxy is to describe workflows. Workflows define a standard set of interconnected analyses that can all be run together rather than step by step. An example of such connections between tools can be described simply using the Workflow Canvas is shown in Figure 29. Workflows help ensure that the same procedure is used each time when an analysis requires repetitive use of that procedure across different cohorts.

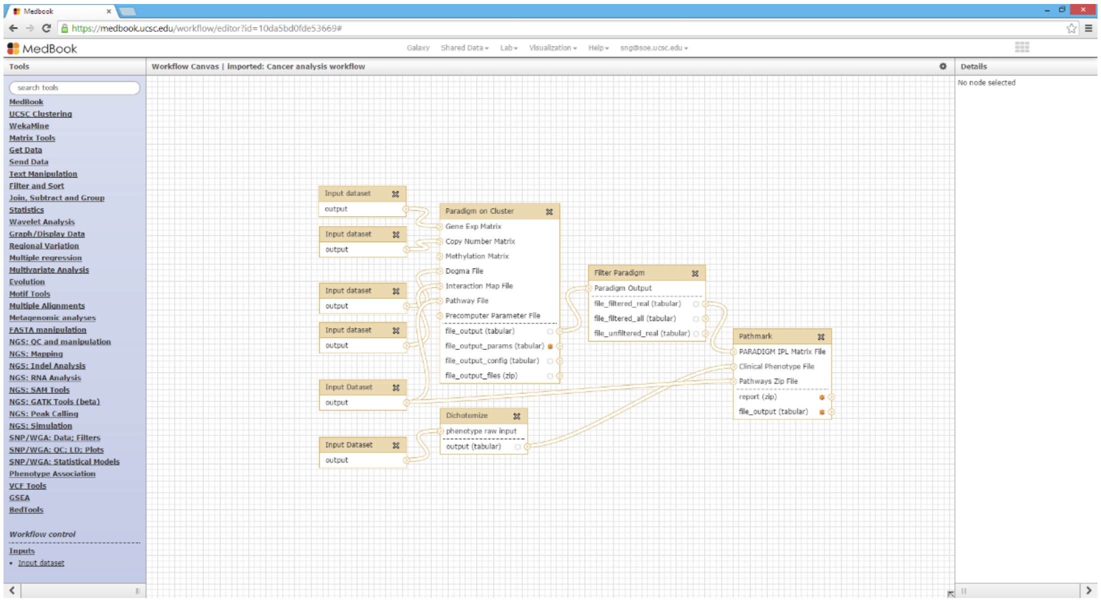


Figure 29 Example workflow from running *PARADIGM* to identifying *PATHMARK* pathway markers.

5.3 Future Challenges

While Galaxy provides an invaluable framework for running computational methods in a reproducible and accessible manner, there are shortcomings that will need to be addressed in order for fluid sharing of tools. I will discuss a few of those challenges here and provide discussion for how they can be improved upon in the future.

5.3.1 Shipping Modules with Docker

Once computation tools have been successfully installed on a Galaxy server, it is relatively easy for collaborators to use these tools. However, managing the different dependencies for each module can quickly lead to broken tools clogging up analysis pipelines. Without a simple way to manage dependencies, installing tools into Galaxy can sometimes be more difficult than getting them to run on the command-line. Docker serves as a potential solution by offering a lightweight virtual container that can be built to describe the dependencies necessary by any tools deployed in Galaxy [Merkel 2014]. This would reduce the overhead on the users by incorporating a Dockerfile containing the dependencies compiled by the developer to ship alongside any developed computational tools.

5.3.2 Enforcing File Formats

Yet another roadblock is the lack of standardized file formats between computational tools. Many tools have come to expect different assumptions about the formatting of the data. Issues as simple as expecting “NA” for missing values versus “Nan” can lead to clogs in pipelines. Strict enforcement of data types is needed to identify that at each step of the pipeline the expected results are observed and no “leaks” are allowed to propagate through the pipeline resulting in inaccurate findings. The danger of propagating errors through analyses grows with the increasing complexity of our workflows, so the necessary caution and overhead will be well worth the effort in the long run. Tools should use standardized formats when possible, but use adaptors between formats should a specialized format be necessary. This would also ensure maximal compatibility between tools.

5.4 Summary

Generating reproducible scientific results is a critical issue that needs to be addressed given the increasing number of publications now being retracted because of either scientific error or fraud. The former can be addressed in part by better bookkeeping to ensure that analytical results become reproducible after a paper is published. This is difficult to achieve without a common framework and guidelines that must be followed. Galaxy has been growing in popularity for integrating biological computational tools with a suite of basic tools and visualizations.

The toolshed concept also allows tools to be portable and removes the necessity to understand UNIX in order to run computational tools. This allows for tools to be more accessible to a wider user base. History and version control ensure that analyses run within Galaxy are reproducible. Our group has been using Galaxy to interface with collaborators and allow them to run our tools. In addition to developing novel tools for pathway-based analysis of genomic data in cancer, I have made my methods available to run through Galaxy.

References

[Adzhubei 2010] Adzhubei I.A., *et al*. (2010) A method and server for predicting damaging missense mutations. *Nature Methods*, **7**, 248-249.

[Anders 2010] Anders S., Wolfgang H. (2010) Differential expression analysis for sequence count data. *Genome Biology*, **11**, R106.

[Baker 2013] Baker M. (2013) Big biology: The 'omes puzzle. *Nature*, **494**, 416-419.

[Blankenberg 2014] Blankenberg D., *et al*. (2014) Dissemination of scientific software with Galaxy ToolShed. *Genome Biology*, **15**(2), 403.

[Benz 1982] Benz C., Tillis T., Tattelman E., Cadman E. (1982) Optimal schedule of methotrexate and 5-fluorouracil in human breast cancer. *Cancer Research*, **42**, 2081-2086.

[Boekel 2015] Boekel J., *et al*. (2015) Multi-omic data analysis using Galaxy. Nature *Biotechnology*, **33**, 137-139,

[Cairns 1975] Cairns J. (1975) Mutation selection and the natural history of cancer. *Nature Review*, **255**, 197-200.

[Califano 2011] Califano A., Butte A., Friend S., Ideker T., Schadt E.E. (2011) Integrative Network-based Association Studies: Leveraging cell regulatory models in the post-GWAS era. *Nature*, **713**, 1-22.

[Chin 2011] Chin L., Hahn W.C.. Getz G., Meyerson M. (2011) Making sense of cancer genomics data. *Genes & Development*, **25**(6), 534-555.

[Ciriello 2012] Ciriello G., Cerami E., Sander C., Schultz N. (2012) Mutual exclusivity analysis identifies oncogenic network modules. *Genome Research*, **22**(2), 398-406.

[Chu 2002] Chu G., Narasimhan B., Tibshirani R., Tusher V. (2002) Significance analysis of microarrays (SAM) software. *Nature*, **5**, 436-442.

[Collisson 2012] Collisson, E.A., Cho R.J., Gray J.W. (2012) What are we learning from the cancer genome? *Nature Review*, **9**(11), 621–630.

[Dankort 2007] Dankort D., Filenova E., Collado M., Serrano M., Jones K., McMahon M. (2007) A new mouse model to explore the initiation, progression, and therapy of BRAFV600E-induced lung tumors. *Genes & Development*, **21**(4), 379-384.

[Dees 2012] Dees N.D., Zhang Q., Kandoth C., Wendl M.C., Schierding W., Koboldt D.C., *et al*. (2012) MuSiC: Identifying mutational significance in cancer genomes. *Genome Research*, **22**(8), 1589–1598.

[Ellis 2012] Ellis M.J., Ding L., Shen D., Luo J., Suman V.J., Wallis J.W., Van Tine B.A., Hoog J., Goiffon R.J., Goldstein T.C., Ng S., *et al*. (2012) Whole-genome analysis informs breast cancer response to aromatase inhibition. *Nature*, **486**, 353‐360.

[Forbes 2008] Forbes S.A., *et al*. (2008) The catalogue of somatic mutations in cancer (COSMIC). *Current Protocols in Human Genetics*, **57**, 10.11.10-10.11.26.

[Friedman 2004] Friedman N. (2004) Inferring cellular networks using probabilistic graphicial models. *Science*, **303**(5659), 799-805.

[Getz 2007] Getz G., Hofling H., Mesirov J.P., Golub T.R., Meyerson M., Tibshirani R., Lander E.S. (2007) Comment on 'The consensus coding sequences of human breast and colorectal cancers.' *Science*, **317**, 1500.

[Golub 1999] Golub T.R., *et al*. (1999) Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. *Science*, **286**, 531-537.

[Goecks 2010] Goecks J., Nekrutenko A., Taylor J., Galaxy Team. (2010) Galaxy: a comprehensive approach for supporting accessible, reproducible, and transparent computational research in the life sciences. *Genome biology*, **11**(8), R86.

[Gonzalez-Perez 2011] Gonzalez-Perez A., Lopez-Bigas N. (2011) Improving the Assessment of the Outcome of Nonsynonymous SNVs with a Consensus Deleteriousness Score, Condel. *American Journal of Human Genetics*, **88**(4), 440-449.

[Gonzalez-Perez 2012] Gonzalez-Perez A., Lopez-Bigas N. (2012) Functional impact bias reveals cancer drivers. *Nucleic Acids Research*, **40**(21), e169.

[Green 2011] Green E.D., *et al*. (2011) Charting a course for genomic medicine from base pairs to bedside. *Nature Perspectives*, ***470***(7333), 204–213. doi:10.1038/nature09764

[Hanahan 2000] Hanahan D., Weinberg R.A.. (2000) The Hallmarks of Cancer. *Cell*, **100**(1), 57-70.

[Heiser 2011] Heiser L.M., Sadanandam A., Kuo W., Benz S.C., Goldstein T.C., Ng S., *et al*. (2011) Subtype and pathway specific responses to anti-cancer compounds in breast cancer. *Proceedings of the National Academy of Sciences*, **109**(8), 2724-2729.

[Hoadley 2014] Hoadley K., Yau C., Wolf D.M., Cherniack A.D., Tamborero D., Ng S., *et al*. (2014) Multiplatform Analysis of 12 Cancer Types Reveals Molecular Classification within and across Tissues of Origin. *Cell*, **158**(4), 929-944.

[Kent 2002] Kent W.J., Sugnet C.W., Furey T.S., Roskin K.M., Pringle T.H., Zahler A.M., Haussler D. (2002) The human genome browser at UCSC. *Genome Research*, **12**(6), 996-1006.

[Kent 2003] Kent W.J., Baertsch R., Hinrichs A., Miller W., Haussler D. (2002) Evolution's cauldron: duplication, deletion, and rearrangement in the mouse and human genomes. *Proceedings of the National Academy of Sciences*, **100**(20), 11484-11489.

[Kumar 2009] Kumar P., Henikoff S., Ng P.C. (2009) Predicting the effects of coding non-synonymous variants on protein function using the SIFT algorithm. *Nature Protocols*, **4**, 1073-1081.

[Leiserson 2013] Leiserson M.D.M., Blokh D., Sharan R. Raphael B.J. (2013) Simultaneous Identification of Multiple Driver Pathways in Cancer. *PLOS Computational Biology*.

[Lim 2009] Lim W.K., Lyashenko E., Califano A. (2009) Master regulators used as breast cancer metastasis classifier. *Pacific Symposium on Biocomputing*, 504-515.

[Lopes 2010] Lopes C.T., Franz M., Kazi F., Donaldson S.L., Morris Q., Bader G.D. (2010) Cytoscape Web: an interactive web-based network browser. *Bioinformatics*, **26**(18), 2347-2348.

[Matthews 2009] Matthews L., Gopinath G., *et al*. (2009) Reactome knowledgebase of human biological pathways and processes. *Nucleic Acids Research*, **37**, D619-D622.

[Merkel 2014] Merkel, D. (2014) Docker: Lightweight Linux Containers for Consistent Development and Deployment. *Linux Journal*.

[McFarland 2013] McFarland C.D., *et al*. (2013) Impact of deleterious passenger mutations on cancer progression. *Proceedings of the National Academy of Sciences*, **110**(8), 2910–2915.

[Mooij 2010] Mooij J.M.. (2010) libDAI: A Free and Open Source C++ Library for Discrete Approximate Inference in Graphical Models. *Journal of Machine Learning Research*, **11**, 2169-2173.

[Mooney 2011] Mooney S.D., *et al*. (2011) Bioinformatic tools for identifying disease gene and SNP candidates. *Methods Molecular Biology*, **628**, 307-319.

[Ng 2003] Ng P.C., Henikoff S. (2003) SIFT: Predicting amino acid changes that affect protein function. *Nucleic Acids Research*, **31**(13), 3812-3814.

[Ng 2011] Ng S., Collisson E.A., Sokolov A., Goldstein T., Gonzalez-Perez A., Lopez-Bigas N., *et al*. (2012) PARADIGM-SHIFT predicts the function of mutations in multiple cancers using pathway impact analysis. *Bioinformatics*, ***28***(18), i640–i646.

[Nicholson 2001] Nicholson R.I., Gee J.M., Harper M.E. (2001) EGFR and cancer prognosis. *European Journal of Cancer*, **37**, 9-15.

[Nishimura 2001] Nishimura D. (2001) BioCarta. *Biotech Software & Internet Report*, **2**(3), 117-120.

[Reva 2007] Reva B., Antipin Y., Sander C. (2007) Determinants of protein function revealed by combinatorial entropy optimization. *Genome Biology*, **8**(11), R232.

[Reva 2011] Reva B., Antipin Y., Sander C. (2011) Predicting the functional impact of protein mutations: application to cancer genomics. *Nucleic Acids Research*, **39**(17), e118.

[Ritchie 2015] Ritchie, M. E. (2015) *limma* powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic Acids Research*, **43**(7), e47.

[Robinson 2010] Robinson M.D., McCarthy D.J., Smyth G.K. (2010) edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics*, **26**(1), 139-140.

[Sandve 2013] Sandve G.K., Nekrutenko A., Taylor J., Hovig E. (2013) Ten Simple Rules for Reproducible Computational Research. *PLOS Computational Biology*.

[Schaefer 2009] Schaefer C.F., *et al*. (2009) PID: the Pathway Interaction Database. *Nucleic Acids Research*, **37**, D674-679.

[Shannon 2003] Shannon P., *et al*. (2003) Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome Research*, **13**(11), 2498-2504.

[Sloggett 2013] Sloggett C., Goonasekera N., Afgan E. (2013) BioBlend: automating pipeline analyses within Galaxy and CloudMan. *Bioinformatics*, **29**(13), 1685-1686.

[Steen 2010] Steen R.G. (2010) Retractions in the scientific literature: is the incidence of research fraud increasing? *Journal of Medical Ethics*.

[Subramanian 2005] Subramanian, A. *et al*. (2005) Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles. *Proceedings of the National Academy of Sciences*, **102**(43), 15545-15550.

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

[TCGA 2012] TCGA *et al*. (2012) Comprehensive molecular characterization of human colon and rectal cancer. *Nature*, **487**(7407), 330–337.

[TCGA 2012] TCGA *et al*. (2012) Comprehensive genomic characterization of squamous cell lung cancers. *Nature*, **489**, 519-525.

[TCGA 2012] TCGA *et al*. (2012) Comprehensive molecular portraits of human breast tumours. *Nature*, **490**, 61-70.

[TCGA 2013] TCGA *et al*. (2013) Integrated genomic characterization of endometrial carcinoma. *Nature*, **497**, 67-73.

[TCGA 2013] TCGA *et al*. (2013) Comprehensive molecular characterization of clear cell renal cell carcinoma. *Nature*, **499**, 43-49.

[TCGA 2014] TCGA *et al*. (2014) Comprehensive molecular profiling of lung adenocarcinoma. *Nature*, **511**, 543-550.

[TCGA 2014] TCGA *et al*. (2014) Comprehensive molecular characterization of gastric adenocarcinoma. *Nature*, **513**, 202-209.

[TCGA] TCGA *et al*. Melanoma Marker Paper. *In submission*.

[TCGA] TCGA *et al*. Lower Grade Glioma Marker Paper. *In submission*.

[Vandin 2012] Vandin F., Upfal E., Raphael B.J. (2012) De novo discovery of mutated driver pathways in cancer. *Genome Research*, **22**(2), 375-385.

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

[Vaske 2010] Vaske C.J., Benz S.C., Sanborn J.Z., Earl D., Szeto C., Zhu J., Haussler D., Stuart J.M. (2010) Inference of patient-specific pathway activities from multi-dimensional cancer genomics data using PARADIGM. *Bioinformatics*, **26**(12), i237.

[Wang 2015] Wang, X *et al*. (2015) Silencing of Long Noncoding RNA MALAT1 by miR-101 and miR-217 Inhibits Proliferation, Migration, and Invasion of Esophageal Squamous Cell Carcinoma Cells. *Journal of Biological Chemistry*, **290**(7), 3925-3935.

[Wong 2013] Wong C.K., Vaske C.J., Ng S., Sanborn J.Z., Benz S.C., Haussler D., Stuart J.M. (2013) The UCSC Interaction Browser: multidimensional data views in pathway context. *Nucleic Acids Research*, **41**, W218-W224.

[Wu 2010] Wu G., Feng X., Stein L. (2010) A human functional protein interaction network and its application to cancer data analysis. *Genome Biology*, **11**, R53.

[Zhu 2009] Zhu J., Sanborn J.Z., Benz S., Szeto C., *et al*. (2009) The UCSC Cancer Genomics Browser. *Nature Methods*, **6**, 239-240.

[Ziogas 2009] Ziogas D., Roukos D.H. (2009) Genetics and Personal Genomics for Personalized Breast Cancer Surgery: Progress and Challenges in Research and Clinical Practice. *Annals of Surgical Oncology*, **16**(7), 1771-1782.