Signaling Pathway Deregulation: Identification Through Genomic Aberrations

And Verification Through Genomic Activity

by

Robert Trevino

A Thesis Presented in Partial Fulfillment
of the Requirements for the Degree
Master of Science

Approved July 2011 by the
Graduate Supervisory Committee:

Seungchan Kim, Chair
Markus Ringner
Huan Liu

ARIZONA STATE UNIVERSITY

August 2011

ABSTRACT

Given the process of tumorigenesis, biological signaling pathways have become of interest in the field of oncology. Many of the regulatory mechanisms that are altered in cancer are directly related to signal transduction and cellular communication. Thus, identifying signaling pathways that have become deregulated may provide useful information to better understanding altered regulatory mechanisms within cancer. Many methods that have been created to measure the distinct activity of signaling pathways have relied strictly upon transcription profiles. With advancements in comparative genomic hybridization techniques, copy number data has become extremely useful in providing valuable information pertaining to the genomic landscape of cancer. The purpose of this thesis is to develop a methodology that incorporates both gene expression and copy number data to identify signaling pathways that have become deregulated in cancer. The central idea is that copy number data may significantly assist in identifying signaling pathway deregulation by justifying the aberrant activity being measured in gene expression profiles. This method was then applied to four different subtypes of breast cancer resulting in the identification of signaling pathways associated with distinct functionalities for each of the breast cancer subtypes.

ACKNOWLEDGEMENTS

I am sincerely grateful to Dr. Seunghan Kim and Dr. Markus Ringner for all of their support and guidance in completing this project. I am also eternally grateful to the Fulbright Fellowship that gave me the opportunity to conduct such important research in Sweden-a beautiful and magical place. I would like to also thank the Lois Roth Endowment for giving the opportunity to continue my research in Sweden. This project also couldn't have been completed without the assistance of the Sysbio Lab at Arizona Sate University. I would like to give a special thanks to Michael Verdicchio, Archana Ramesh, and Ina Sen for guiding me along the way. I am also grateful to the Air Force Research Laboratory Mesa Research Site for supporting me throughout my scholastic experience. Finally, I would like to thank God for giving me the wisdom, conviction, and determination in pursuing my dreams. This thesis is dedicated to my father, Roberto Trevino, and beautiful sister, Yvette Trevino.

Contents

Page

LIST OF TABLES vi

LIST OF FIGURES vii

CHAPTER 1

1 INTRODUCTION 1

1.1 Signaling Pathways and Cancer 1

1.2 Current Data Available For Signaling Pathway Analysis 2

1.3 Weighted Signaling Pathway Impact Analysis with GISTIC Genes (WSPIAGG) 3

1.4 Organization Of Thesis 4

2 BACKGROUND 5

2.1 Pathway Analysis 5

2.2 Context-Specific Gene Regulatory Network 6

2.3 Genomic Identification of Significant Targets In Cancer (GISTIC) . . 9

3 WEIGHTED SIGNALING PATHWAY IMPACT ANALYSIS with GISTIC GENES (WSPIAGG) 11

3.1 Original Signaling Pathway Impact Analysis (SPIA) 11

3.2 Improving SPIA 16

Context-Specific Gene Regulatory Networks Improves Quality Of Putative Data 17

Copy Number Data Provides Useful Insight Into Deregulation 20

Combining P-values to determine significance of evidence 28

3.3 Implementation 30

High Throughput Data 31

Identifying Potentially Deregulated Pathways For WSPIAGG Analysis 31

4 DATA ANALYSIS AND RESULTS 32

4.1 Application To Breast Cancer Dataset 32

Chapter Page

Tumor Samples Information 32

Gene Expresion Data 32

Cellular Context Mining On Gene Expression Data 33

Copy Number Data 33

GISTIC Implementation On Copy Number Data 34

Pathway Repositories 34

Pathway Commons 35

Wikipathways 35

KEGG 36

Biomodels 36

4.2 Results 39

Basal Like Analysis 39

Comparison Results 39

Pathways Of Interest 40

Overall Survival Analysis 42

Luminal Complex 43

Comparison Results 43

Pahtways Of Interest 43

Overall Survival Analysis 45

17q12 46

Comparison Results 46

Pathways Of Interest 47

Overall Survival Analysis 47

Luminal Simple 49

Comparison Results 49

Pathways Of Interests 51

Overall Survival Analysis 52

5 DISCUSSION 54

Future Research 55

REFERENCES 57

LIST OF TABLES

Table Page

3.1 Hypergeometric Distribution 12

3.2 Pathway 1 14

3.3 Pathway $1'$ 14

3.4 Fisher's exact test for a set of GISTIC genes found in a given pathway . 21

3.5 Gene Influence Score for each differentially expressed gene in Pathway $1'$ taking into consideration only $maxPF$ 23

3.6 A summary of p-value score generated for each type evidence analyzed 29

4.1 KEGG interactions definition for edge values 38

4.2 Wikipathways interactions definition for edge values 38

4.3 Comparison between WSPIAGG and SPIA in Basal-like subtype 40

4.4 Active pathways in Basal-like subtype 42

4.5 Logrank test for each clustered group in Basal-like subtype 42

4.6 Comparison between WSPIAGG and SPIA in Luminal Complex subtype 44

4.7 Active pathways in Luminal Complex subtype 46

4.8 Logrank test for each clustered group in Luminal Complex subtype . . . 46

4.9 Comparison between WSPIAGG and SPIA in 17q12 subtype 48

4.10 Active pathways in 17q12 subtype 48

4.11 Logrank test for each clustered group in 17q12 subtype 50

4.12 Comparison between WSPIAGG and SPIA in Luminal Simple subtype . 50

4.13 Active pathways in Luminal Simple subtype 51

4.14 Logrank test for each clustered group in Luminal Simple subtype 53

LIST OF FIGURES

Figure Page

2.1 GIST̄IC algorithm overview provided by [1] 10

3.1 Comparison of pathways could yield similar gene enrichment results . . 13

3.2 Edges are strengthened using Context-Specific GRN 17

3.3 (a) Markov Clustering Results [TN study]. (b) Asymmetric Spectral Clustering Results [TN study]. Clustering previous data resulted in identification of contexts enriched with different cancer types. 18

3.4 Comparison of influence on pathway $P1'$ between genes 23

3.5 States pertaining to high throughput data for a given gene plotted in Cartesian coordinate system 25

4.1 Pathway activity as measured by WSPIAGG in Basal-like subtype . . . 41

4.2 Pathway activity as measured by SPIA in Basal-like subtype 41

4.3 Overall survival plots for clustered group 4 in Basal-like subtype 43

4.4 Pathway activity as measured by WSPIAGG in Luminal Complex subtype 45

4.5 Pathway activity as measured by SPIA in Luminal Complex subtype . . 45

4.6 Overall survival plots for clustered group 4 in Luminal Complex subtype . 47

4.7 Pathway activity as measured by WSPIAGG in 17q12 subtype 49

4.8 Pathway activity as measured by SPIA in 17q12 subtype 49

4.9 Overall survival plots for clustered groups in 17q12 subtype 51

4.10 Pathway activity as measured by WSPIAGG in Luminal Simple subtype 52

4.11 Pathway activity as measured by SPIA in Luminal Simple subtype . . . 52

4.12 Overall survival plots for clustered groups in Luminal Simple subtype . . 53

Chapter 1

INTRODUCTION

1.1 Signaling Pathways and Cancer

Signal transduction describe a series of molecular interactions where external cellular signals induce an intracellular response. These responses can influence a cell by creating protein products that induce transcriptional and metabolic behaviors. The specific steps with respect to gene and proteins interactions that occur when signal transduction is initiated can be described using biological signaling pathways. Biological signaling pathways have become of significant interest in oncology due to overwhelming evidence suggesting that tumorigenesis is largely linked to intracellular and intercellular signal degradation and alteration. Hanahan and Weinberg initially identified 6 specific steps that must occur to alter the inherent cellular regulatory mechanisms causing a normal cell to enter into a cancerous state [2]. Out of the six original steps towards tumorigenesis, three were directly related to signal transduction. First, self sufficiency in growth signals is achieved, which is best described as the process by which the autocrine signaling is used to induce self growth without any assistance of signaling from other cells [3] . Second, insensitivity to antigrowth signals is accomplished, which is defined as the breakdown of paracrine signaling by which neighboring cells try to communicate to a rogue cell to cease in growing without success. Finally, evasion of apoptosis represents a culmination of breakdown of all signaling where apoptosis or programmed cell death (pcd) of a rogue cell cannot be initiated internally by the rogue cell or externally by neighboring cells. In addition, recent evidence has been presented suggesting that cancer cells rely on interactions with normal cells in nearby surrounding areas forming a complex micro-environment composed of a mixture of normal and cancerous cells [4]. This has profound implications in that signal transduction networks are not completely destroyed but rather altered to benefit a cancerous environment. Signaling pathway analysis, therefore, provides a starting point to identify and better understand the blueprint of altered and re-wired signal transduction in cancer.

1.2 Current Data Available For Signaling Pathway Analysis

Microarray technology has revolutionized biology by expanding our insight into the genetic inner-workings of a cell. Gene expression profiling has played an especially critical role in understanding cellular processes by simultaneously measuring the activity, also known as expression, of thousands of genes at once. Gene activity can be measured based on the amount of messenger Ribonucleic Acid (mRNA) that is produced or expressed. Strands of complementary Deoxribonucleic Acid (cDNA) are first synthesized from mRNA using enzymes and marked with florescent markers. The cDNA from different target genes are then introduced to a platform called a microarray that contains thousands of known regions or sections of DNA called probes. Using hybridization techniques, binding occurs between cDNA and the probes. Using the florescent markers, image analysis is then done where the intesity of the spots of each marker are converted into an expression value. Many microarrays can be combined into an expression matrix, which is then transformed into the famous heat maps that are commonly used in bioinformatics. This technology has proven to be extremely useful in the analysis of signaling pathways and cancer as a whole.

A much newer method that has proven to be just as pertinent in cancer research is that of array-based Comparative Genomic Hybridization aCGH, which focuses on the genomic landscape of a cell. Each cell in the human body contains the entire genome, with the exception of sex cells which are beyond the scope of this thesis. In order to maintain a healthy state, a cell must keep intact as much of it's entire genome as possible and has regulatory mechanisms in place to do so. If these regulatory mechanisms are compromised, significant aberrations and mutations may occur in a cell's genome. Tumorigenesis is directly linked to these mutations that occur on the genome of a cell. These mutations allow a cell to bypass many biological safeguards that were intended to prevent disruption of normal cellular functionality causing it to enter a highly proliferative, uncontrolled state. Cancer cells are unique in that they reproduce defying normal restraints on cells, metastasizing and colonizing in other biological environments normally reserved for other cell types [3]. Thus, it is quite conceivable to ascertain the health of a cell by analyzing its genomic structure. The method employed by aCGH is based on the same principles as gene expression profiling. However, there are some significant differences worth noting. First, a test cell and a reference cell's DNA are cleaved using different enzymes into smaller portions depending on the size of the DNA sequence or window that is selected. The smaller the window selected the larger and more specific the resulting data set is. Next, each cell's DNA is labeled differently to discern one from the other using florescent markers. They are then hybridized to thousands of probes and the resultant intensity ratio is used to measure copy number alterations in the test cell. The aCGH technique has proved beneficial in providing the data needed to develop methods for better understanding the genomic structure of different types of cancer such as breast cancer and glioma [5, 1]. Copy number data has also been important in signaling pathway research, though, to a much less extant as its gene expression counterpart.

1.3 Weighted Signaling Pathway Impact Analysis with GISTIC Genes (WSPIAGG)

Inspired by two robust algorithms, Signaling Pathway Impact Analysis (SPIA) and Genomic Identification of Significant Targets In Cancer (GISTIC) method, Weighted Signaling Pathway Impact Analysis with GISTIC Genes (WSPIAGG) was developed for signaling pathway analysis. This method incorporates the use of gene expression data with copy number data to better understand the impact that gene mutations have on expression data impacting signaling pathways. In addition, interactions in each pathway are verified through the use of cellular context mining (ccm), a powerful tool capable of recognizing gene regulatory relationships in gene expression data. The overall purpose of combining these methods was to give more credence to pathways that contained mutated genes where sufficient evidence exists that they may be influencing the activity of other genes within a given pathway.

1.4 Organization Of Thesis

The framework and origins of the WSPIAGG method is provided in the background. WSPIAGG is formally defined in chapter 3. Since often times methodologies become too complex and incapable of being implemented, chapter 3 also describes its implementation. Chapter 4 describes the real-world application of WSPIAGG by applying it to breast cancer data sets. The method is then compared against the original SPIA methodology analyzing performance of identifying pathway activity as well as the average score given across different tumor samples. Chapter 5 provides an explanation for the differing results when comparing the two methodologies. Given that the expansion of WSPIAGG to use other data and tools is quite feasible in the not too distant future, chapter 5 also describes future works.

Chapter 2

BACKGROUND

2.1 Pathway Analysis

Several methods for analyzing signaling pathways using transcription profiles have been developed. Original methods focused on overrepresentation analysis (ORA) of differentially expressed genes within a pathway [6]. However, as understanding of signaling pathways have increased, the methodology to determine signaling pathway deregulation has improved.

Some methods focused on determining the activity of signaling pathways in gene expression profiles based on the gene activity of target genes. If a pathway is deregulated, transcription factors influenced in a signaling pathway will affect the expression of its target. This was suggested given that signaling pathways may be deregulated and not display transcriptional activity of it's member genes. Breslin et al. demonstrated that analysis of downstream target genes within a sample was a viable option for identifying signaling pathway deregulation [7]. Liu and Ringner further proposed analyzing transcription factors that mediated signaling pathways and using corresponding cis-regulatory motifs to identify potential genes that may show activity if a given pathway were to be deregulated [8]. Both cases used knowledge of transcription factors and target genes to determine deregulation of signaling pathways. However, gene regulatory networks within a signaling pathway were not taken into consideration in both methods when identifying deregulation.

Tarca et al. developed a method that incorporated the graph structure of a pathway with the transcriptional activity of member genes to assess the full impact of differentially expressed genes [9]. It was a significant improvement over other pathway analysis methods that relied on ORA only or didn't take into consideration inherent gene regulatory networks when determining pathway deregulation. Moreover, it took advantage of pathway information from KEGG Pathway database to validate the different interactions types between genes. One of the core strengths within this method was that it did not try to replace ORA analysis but rather complement it through the introduction of a novel algorithm known as the pathway perturbation factor. This perturbation factor measured how a gene might propagate its influence on genes that are downstream of it in a given pathway. Another core strength was its robustness and ability to allow for implementation of strengthening the known interaction between two genes in a pathway. The WSPIAGG method was built upon this method because of these core strengths.

2.2 Context-Specific Gene Regulatory Network

Gene expression data has proven useful in giving an overall picture of a tumor sample's gene activity, but it has been quite difficult to discern how this activity explains cellular states and the corresponding genetic interactions. In an effort to better explain the interaction of genes being regulated in the different states or contexts of a cell, Doughtery et al introduced a mathematical model for describing contextual gene regulation [10]. This mathematical model assumes that within a specific context there are M sets, $G1$, $G2$, ..., GM, of driver genes and m corresponding sets, $S1$, $S2$,..,Sm, of driven genes and that for each driven set Sj there is a driver set Gj that is governing the behavior of genes in Sj. The significance of this fact is self evident when environmental factors that cause mutations of a gene may correspondingly change other gene expressions. Transcriptional changes could impact normal regulatory mechanisms and, thus, change the overall state of a cell.

A cell enters into a cancerous state when normal regulatory mechanisms have changed and adjusted to environmental factors to provide proliferative signals and usurp inherent biological safeguards intended to prevent abnormal cells. Cancerous cells may eventually create micro-environments that retain a complex, consistent, and reliable regulatory machinery that is required for a cell to survive and proliferate. In taking advantage of potentially consistent transcriptional behavior within a cellular state, Kim et al. developed an algorithm that uses gene expression data to identify the relationship between sets of genes within a specific biological state known as a context motif [11]. A context motif can be thought of as the state of a cell defined by the transcriptional activity of a set of genes regulating another set of genes within a subset of samples that share some phenotypic attributes. The two key statistical parameters that are used to determine the activity and regulation of genes within a specific context motif are interference and cross talk. The interference of a gene in a context motif is defined as "the extent to which latent variables (external controls sensitive but not specific context motif) interfere with regulatory signals from a master gene, G_j" [11].

$$\delta_k^j = 1 - Pr(g_k = ON|C = c_j) \tag{2.1}$$

The crosstalk of a gene is defined as the probability that the gene, g_k is being regulated (by external control), when the cellular context is not c_j.

$$\eta_k^j = Pr(g_k = ON|C \neq c_j) \tag{2.2}$$

Context motifs hold two important graph structure properties that are worth noting. First, the driver-driven relationship between sets of genes within a context motif form a directed graph. Second, a gene may be a driver gene in one context motif while simultaneously being a driven gene in another displaying an overlapping community structure that is often seen in nature [12, 13]. Noting the inherent graph properties of context motifs identified in cellular context mining, Sen et al. developed a method of formally constructing context-specific gene regulatory networks from context motifs [14]. Through the combination of various context motifs that had overlapping genes and taking advantage of the directionality of interaction, an interesting graph structure emerged representing a community of genes regulating one another. Thus, the overall graph structure was called a context-specific -gene regulatory network (GRN). The context-specific GRNs were then grouped based on the sparsity of edges seen by the human eye. The different groups identified were classified as contexts since they represented overlapping gene activity between groups of context motifs. A sample association score was then developed to determine, as its name implies, which context a tumor sample was closely associated with.

$$SAS(s,C) = \sqrt[m]{\prod_{i=1}^{m} f_i(s)} \text{ where } f_i(s) = \begin{cases} k_i/N, \ s \in C_i \\ 1, \ \text{otherwise} \end{cases} \tag{2.3}$$

where k_i is the number of samples within a context C_i and N is the total number of samples in the gene expression data. Sen et al. demonstrated that given a mixture of tumor samples pertaining to different cancer types, the resultant contexts formed from the different tumor samples analyzed were statistically enriched with the different types of cancer [14].

Ramesh et al. further investigated the graph structure of context-specific GRNs by comparing the contexts that resulted from applying two different clustering algorithms [15]. Traditionally, bottom-up or agglomerative approaches in hierarchical clustering have been applied to transcription profiles to identify groups of significantly important genes. However, this approach has two limitations that would prohibit use on context-specific GRNs. First, the time and space complexity for m data points can reach as high as $O(m^2 \log m)$ and $O(m^2)$, respectively [16]. Given the size of nodes or data points and the density of edges, it would not be efficient or even feasible in applying to context-specific GRNs. Second, given its bottom up nature, hierarchical clustering used in transcription profiles lacks any global objective [16] which is paramount in context-specific GRNs given the relationship between different context motifs. Thus, two clustering algorithms were selected that implemented a top-down or divisive approach while taking into consideration the global view of the data. Markov Clustering Algorithm (MCL) [17] and spectral clustering

[18] were applied to context-specific GRNs to verify if any significant biological inferences could be gleaned in and efficient manner. Contexts obtained from spectral clustering and MCL clustering were compared on a number of different attributes such as connectivity density within and between clusters. These different contexts were also analyzed for enrichment of different cancer types that occurred with statistical significance. The significance of this study implicates that transcription profiles from different types of cancer may be grouped together from a top-down approach and still yield significant results when implemented with context-specific GRNs. MCL clustering provided a much more robust method given that one does not need to identify the number of cluster beforehand. Moreover, MCL performed comparably well if not better than spectral clustering with respect to coverage and performance values [15]. Thus, the use of contexts clustered using MCL were the cornerstone of inferred interaction data in WSPIAGG.

2.3 Genomic Identification of Significant Targets In Cancer (GISTIC)

With respect to genomic structure, the GISTIC method [1] proved to be efficient and simple in its application and extremely useful in taking chromosomal data and translating it into pertinent information of cancer . The method first identifies different areas of chromosomal aberrations across a set tumor samples. The method then assigns a G score to these previously identified areas based on the total magnitude of aberrations, in essence, summing them up. These aberrations are then permuted in each sample across the genome and the G score is recalculated to determine the probability of finding the observed G score by random chance. Those aberrations with high amplitude consistent across the samples are considered significant in the respective type of cancer that the tumor samples are associated with. Genes that are found within regions identified by the GISTIC method are termed GISTIC genes and are used for further analysis in WSPIAGG.

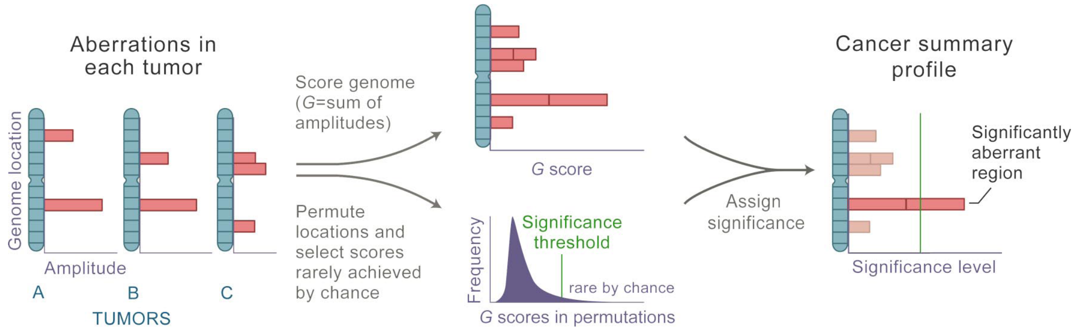


Figure 2.1: GISTIC algorithm overview provided by [1]

This method has assisted in identifying chromosomal regions of significance in both glioma and breast cancer [1, 5]. Moreover, in the particular case of breast cancer it was used to assist in the identification of six subtypes of breast that shared similar clinical characteristics [5].

Chapter 3

WEIGHTED SIGNALING PATHWAY IMPACT ANALYSIS with GISTIC GENES (WSPIAGG)

The WSPIAGG scoring method is a combination of three different methods that have proven useful in cancer research. It improves the original Signaling Pathway Impact Analysis (SPIA) [9] in two distinct ways. First, it incorporates data from Cellular Context Mining (CCM) to strengthen gene interactions in a pathway where evidence exists. The original SPIA method allowed for specifying the strength of the interactions given that a fixed value be defined for the different types of interactions in KEGG database [19]. The use of CCM is much more dynamic, strengthening interactions based on gene-pair interactions inferred through real world evidence in the form of gene expressions profiles. Second, genes identified through the use of the GISTIC method are used in determining the significance of any perturbation measured. The GISTIC method introduce the ability to identify specific genes that may be of importance in signaling pathways where aberrant activity has been measured. It incorporates a necessary component of genomic structure in signaling pathway analysis. The original SPIA method is also quite efficient in calculating the perturbation scores as will be demonstrated using vectors and matrices. The WSPIAGG method will build upon this efficiency in modifying the original methodology as well as in computing the newly introduced values.

3.1 Original Signaling Pathway Impact Analysis (SPIA)

Signaling Pathway impact analysis (SPIA) combines two independent forms of evidence in the analysis of signaling pathways. The first is the traditional form of overrepresentation analysis (ORA), obtaining the probability of finding a set of differentially expressed genes within a pathway. ORA is well known for its simplicity and reliability and is traditionally used in extracting useful information from gene expression profiles. ORA is done by finding the number of pathway genes that

	Member Set	Non-member Set	total
Differentially Expressed	d	$m-d$	m
Normally Expressed	$n-d$	$N+d-n-m$	$N-m$
total	n	$N-n$	N

Table 3.1: Hypergeometric Distribution

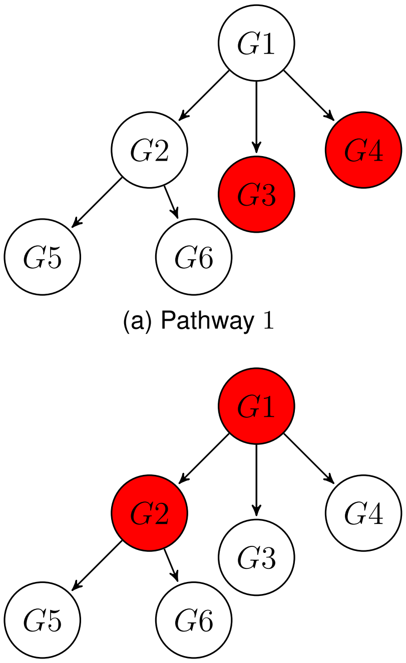
are differentially expressed versus the number of pathway genes with no differential expression given the total pathway genes found in the gene expression profile. The probability P_{NDE} of finding a given number of differentially expressed genes is calculated using the hypergeometric distribution.

$$P(M=m) = \frac{\binom{m}{d} \times \binom{N-m}{n-d}}{\binom{N}{n}} \tag{3.1}$$

In an ideal scenario, expression data would exist for every gene in a pathway. However, at this point no pathway repository contains this amount of information. Therefore calculating the enrichment of differentially expressed genes, DE_g, within the confines of a set of pathway genes, P_g , must be done by taking the intersection, $P_g \cap DE_g$, of the two to represent d while N should equal the intersection, $T_g \cap P_g$, of all genes with gene expression data, T_g, and pathway genes, P_g.

The second form of evidence used in SPIA is referred to as the perturbation analysis. This analysis exploits the graph structure created by gene member interactions to determine the full impact of a differentially expressed gene in a signaling pathway. This is imperative to help mitigate some of the short-comings that simple overrepresentation analysis has. A useful demonstration provided by Tarca et al [9] compared two hypothetical pathways that have the same number of differentially expressed genes but differ in which genes are differentially expressed.

As demonstrated by the comparison examples, the two pathways will have the same enrichment p-value associated with them. However, given the graph



(a) Pathway 1

(b) Pathway $1'$

Figure 3.1: Comparison of pathways could yield similar gene enrichment results

structure of Pathway $1'$, it is much more likely of extracting useful information from gene interactions than Pathway 1. This is because genes that are further upstream of other genes and are differentially expressed have a much higher probability of influencing downstream genes. Perturbation analysis takes advantage of the graph structure by measuring the accumulated perturbation within a pathway. The amount of perturbation is measured at a single gene in a pathway using a perturbation factor $PF(g)$.

$$PF(g) = \Delta E(g_i) + \sum_{j=1}^{n} \beta_{ij} \times \frac{PF(g_j)}{N_{ds}(g_j)} \tag{3.2}$$

where $\Delta E(g_i)$ represents the signed normalized measured expression change of gene g_i in a sample.

The expression of gene g_i is added to the sum of perturbation factors of directly upstream genes g_j, normalized by the number of downstream neighbors that each g_j has, $N_{ds}(g_j)$. The strength of the interaction between genes g_i and g_j is

Gene	ΔE	PF	Acc
G1	0	0	0
G2	0	0	0
G3	1.5	1.5	0
G4	2	2	0
G5	0	0	0
G6	0	0	0
Total Acc			0

Table 3.2: Pathway 1

Gene	ΔE	PF	Acc
G1	1.5	1.5	0
G2	2	2.5	0.5
G3	0	1.25	1.25
G4	0	1.25	1.25
G5	0	.5	.5
G6	0	.5	.5
Total Acc			4

Table 3.3: Pathway $1'$

quantified through the absolute value of β_{ij}, while the directionality is represented by assigning β_{ij} a value of 1 corresponding to activation and -1 corresponding to inhibition. Thus, a resultant matrix, β, is used to represent the strength and directionality of interaction from gene g_j to g_i .

$$\beta = \begin{pmatrix} \beta_{11} & \beta_{12} & \cdots & \beta_{1j} \\ \beta_{21} & \beta_{22} & \cdots & \beta_{2j} \\ \vdots & \vdots & \ddots & \vdots \\ \beta_{i1} & \beta_{i2} & \cdots & \beta_{ij} \end{pmatrix}$$

The perturbation factor rewards differentially expressed genes that have the potential to influence other genes in a pathway making use of a pathway's graph structure.

Tarca et al. demonstrated that if similar expression values were assigned to each of the differentially expressed genes within Pathway 1 and Pathway $1'$ the total perturbation accumulation would be significantly higher in Pathway $1'$ [9] .

The importance of the position of each differentially expressed genes in a pathway K is, therefore, captured and given a quantifiable score. In order to ensure that disconnected genes are not considered in perturbation analysis and that the gene expression of gene g_i is not double counted in ORA and perturbation analysis, the gene expression of gene g_i, $\Delta E(g_i)$, is subtracted from the perturbation factor measured at g_i. This value is considered an accumulated perturbation measurement, $Acc(g_i)$, at level g_i .

$$Acc(g_i) = PF(g_i) - \Delta E(g_i) \tag{3.3}$$

An Acc vector containing the accumulated perturbation for each gene in a pathway can efficiently be calculated by setting B_{ij} to equal β_{ij} divided by the number genes that are downstream, $N_{ds}(g_j)$,of gene g_j.

$$B = \begin{pmatrix} \frac{\beta_{11}}{N_{ds(g_1)}} & \frac{\beta_{12}}{N_{ds(g_2)}} & \cdots & \frac{\beta_{1j}}{N_{ds(g_j)}} \\ \frac{\beta_{21}}{N_{ds(g_1)}} & \frac{\beta_{22}}{N_{ds(g_2)}} & \cdots & \frac{\beta_{2j}}{N_{ds(g_j)}} \\ \cdots & \cdots & \ddots & \cdots \\ \frac{\beta_{n1}}{N_{ds(g_1)}} & \frac{\beta_{n2}}{N_{ds(g_2)}} & \cdots & \frac{\beta_{nj}}{N_{ds(g_j)}} \end{pmatrix}$$

Applying the following equation will yield the resultant accumulation vector.

$$Acc = B \cdot (I - B)^{-1} \cdot \Delta E \tag{3.4}$$

where ΔE is the vector of all gene expression values in a pathway.

$$\Delta E = \begin{pmatrix} \Delta E1 \\ \Delta E2 \\ \vdots \\ \Delta En \end{pmatrix}$$

I is simply an $n \times n$ identity matrix where n represents the number of genes in a pathway.

The total accumulated perturbation in a pathway could then be computed by summing the resultant Acc vector.

$$t_A = \sum_{i=1}^{n} Acc \tag{3.5}$$

Tarca et al. demonstrated that higher total accumulated perturbations were less probable in a pathway then lower accumulated perturbations. The probability of finding this score by random chance is calculated using a simple bootstrap technique with replacement. The bootstrap technique can be defined as for each pathway K, a set of $N_{de}(P_i)$ differentially expressed gene IDs intersecting pathway K and the gene set of the complete gene expression profile are selected and a random perturbation accumulation score $T_A(K)$ is re-calculated. This process is conducted N_{ite} times where the larger the number the more accurate the probability will be. The random median perturbation accumulation score T_A is then calculated and subtracted from the random accumulation scores $T_A(K)$ to center the distribution around 0 giving $T_{A,c}(K)$. In addition, the median T_A is subtracted from the observed pathway score to correct for the shift in the null distribution median giving $t_{A,c}$. The probability P_{PERT} is obtained using the following equations.

$$P_{PERT} = \begin{cases} 2 \times \frac{\sum_k I(T_{A,c}(K) \geq t_{A,c}) \text{ if } t_{A,c} \geq 0}{N_{ite}} \\ 2 \times \frac{\sum_k I(T_{A,c}(K) \leq t_{A,c}) \text{ otherwise}}{N_{ite}} \end{cases} \tag{3.6}$$

If $t_{A,c} > 0$ then the pathway is considered activated and if $t_{A,c} < 0$ then the pathway is considered inhibited.

The p-value associated with each of the evidences was than combined using

$$PG = c_i - c_i \cdot ln(c_i)$$

where $c_i = P_{PERT} \cdot P_{NDE}$. This method proved superior to the use of only ORA on a pathway as was demonstrated by Tarca et al. [9].

3.2 Improving SPIA

Building upon the concepts introduced by SPIA, two main areas will be strengthened. The first is the use of transcription profile data in conjunction with cellular

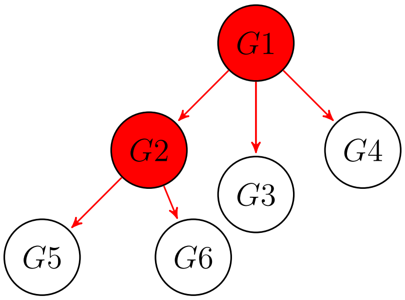


Figure 3.2: Edges are strengthened using Context-Specific GRN

context mining (CCM) from the tumors being analyzed to strengthen the putative interactions in pathways. The second is the introduction of copy number data in assisting with the identification of genes that may be causing a pathway's aberrant activity. Each of these contributions will provide a new level of accuracy in determining pathway deregulation.

Context-Specific Gene Regulatory Networks Improves Quality Of Putative Data

The robustness of the SPIA method is due to to its ability of allowing the strengthening of the putative gene interaction through modification of β_{ij}. Original results generated by Tarca et al. on colorectal cancer datasets used $|\beta = 1|$ in order to minimize the number of model parameters in the research conducted [9] . However, strengthening putative interactions in a given pathway with reliable data obtained from the actual sample expression profiles would be ideal given the nature of cancer. As such, context mining was augmented to support putative interactions in a pathway. Context mining has proven to infer useful insight into driver-driven relationships between genes across different samples of expression data in cancer [11, 14, 15]. Using context mining, different states known as context motifs, CM, consisting of a gene or set of genes that have a high statistical probability of influencing another set of genes can be identified. A context-specific gene regulatory network (GRN) can then be constructed based on the gene overlap between different context-motifs. However, the resultant GRN may be huge making further

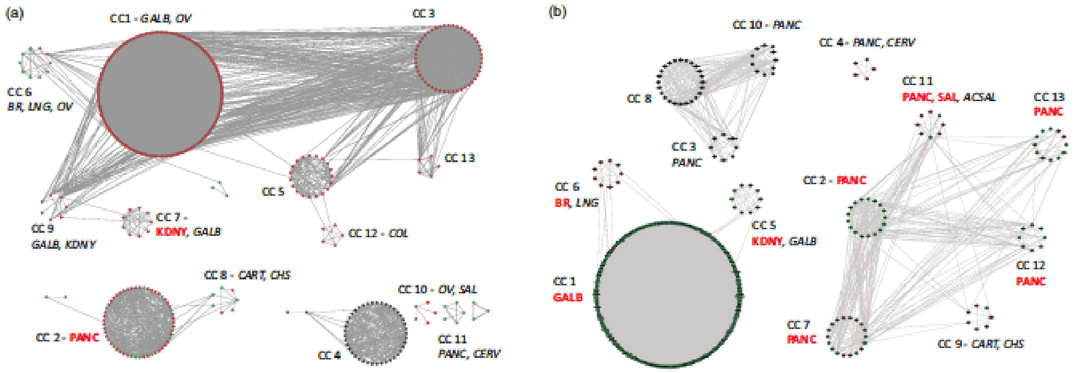


Figure 3.3: (a) Markov Clustering Results [TN study]. (b) Asymmetric Spectral Clustering Results [TN study]. Clustering previous data resulted in identification of contexts enriched with different cancer types.

analysis quite cumbersome if not impossible.

My previous published works of clustering context-specific GRNs into more manageable networks or contexts was shown to be a viable option capable of identifying multiple clusters enriched with different types of cancer [15]. In particular, MCL proved to be useful and efficient in its application with respect to tumor analysis and therefore is used to split the resultant context-specific GRN's.

This allows for the further strengthening of putative gene interactions when analyzing gene expression profiles from samples across different types or even subtypes of cancer. The strengthening of the interaction between two genes is simply done by taking the inverse of the summed number of nodes traversed (hops), $GRNH_{ij}$, from gene g_i to g_j in the context-specific GRN and adding it to the absolute value of the putative interaction then reapplying the original sign value as as shown in equation 3.7.

$$\beta_{ij}^{'} = p_{ij} \times (1 + \frac{1}{GRNH_{ij}}) \tag{3.7}$$

where p_{ij} is the putative interaction value in a pathway between two genes.

$$GRNH_{ij} = \sum_{1}^{h} 1 + \epsilon \tag{3.8}$$

A user-defined error, ϵ, can be added to each hop that is outside the cluster of those enriched by the subtype of the sample expression profiles being analyzed. The hops value, $GRNH_{ij}$, transforms β_{ij} into a stronger $\beta^{'}_{ij}$ dependent upon the composition of a context-specific GRN. Thus, an optimal scenario can be considered as gene g_j having a summed hops value to gene g_i of $GRNH_{ij} = 1$ indicating that strong evidence exists in the expression profile that gene g_j is influencing g_i. This scenario would double the strength of the interaction between gene g_j and gene g_i and therefore increase the accumulated perturbation in the original SPIA method transforming the original perturbation factor into.

$$PF^{'}(g) = \Delta E(g_i) + \sum_{j=1}^{n} \beta^{'}_{ij} \times \frac{PF(g_j)}{N_{ds}(g_j)} \tag{3.9}$$

In order to compute Acc', a matrix $GRNH$ is derived with values representing 1 divided by the number of nodes traversed (hop value), $GRNH_{ij}$, from genes g_j to g_i within the clustered context-specific GRN.

$$GRNH = \begin{pmatrix} \frac{1}{GRNH_{11}} & \frac{1}{GRNH_{12}} & \cdots & \frac{1}{GRNH_{1j}} \\ \frac{1}{GRNH_{21}} & \frac{1}{GRNH_{22}} & \cdots & \frac{1}{GRNH_{2j}} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{1}{GRNH_{i1}} & \frac{1}{GRNH_{i2}} & \cdots & \frac{1}{GRNH_{ij}} \end{pmatrix}$$

Applying the $GRNH$ matrix to the B matrix represents the strengthening of the interactions forming B'.

$$B' = B + (GRNH \times \beta)$$

In order to preserve the same interaction sign, $GRNH$ is first multiplied by the β before adding it to B. The weighted accumulated perturbation vector, Acc', was then derived using the following equation as specified in [9].

$$Acc' = B' \cdot (B' - I)^{-1} \cdot \Delta E$$

The sum of the Acc' vector represents the total modified accumulated perturbation of a pathway $t'_A = \sum Acc'$. Random scoring is done $N_{ite} > 2000$ to keep the same random scoring parameters as those described in the original SPIA [9] for generating P'_{PERT}.

Copy Number Data Provides Useful Insight Into Deregulation

The SPIA method is quite useful in measuring the potential activity of a pathway but does not make an attempt to explain the origins of such activity. To this extent, knowing the molecular structure of genes within a pathway may better explain why such aberrant activity is occurring. Similar to the original SPIA method, two forms of evidence are used in determining the significance of genomic aberrations in a signaling pathway.

The first evidence is the enrichment analysis of a specific type of mutated gene known as a GISTIC gene. These GISTIC genes are derived from areas of chromosomal aberrations identified through the use of the Genomic Identification of Significant Targets In Caner (GISTIC) method. As previously discussed, the GISTIC method has been extremely useful in its application to cancer datasets. Therefore, analyzing pathways for GISTIC gene members could provide crucial evidence for deregulation analysis. Fisher's exact test is used to determine the probability of finding a number of GISTIC genes in a given pathway randomly, identifying these pathways as potentially significant in the cancer subtype being analyzed. Only the intersecting set, $GG \cap Db_g$, of GISTIC genes, GG, and those genes located in the

Fisher's Exact Test For Pathways

	In Pathway	Not In Pathway
GISTIC Genes	a	b
non-GISTIC Gene	c	d

Table 3.4: Fisher's exact test for a set of GISTIC genes found in a given pathway

database the pathway originated from, Db_g, were considered as the total number of GISTIC genes, $a + b$, when calculating the p-value.

The less likely the number of GISTIC genes in a pathway are, the more likelihood that they may play a role in the activity of a pathway if it is deregulated. This supported by the fact that chromosomal aberrations identified using the GISTIC method have proven to be quite useful in classifying subtypes of cancer with similar clinical characteristics [5, 1]. Moreover, many of the GISTIC genes identified within these areas of aberration have shown to be quite interesting with respect to their functional annotations. Therefore, they are used as markers in identifying signaling pathways that may be potentially deregulated. Any pathway that's member genes, P_g, intersect the set of GISTIC genes, GG, are considered to be in a potentially deregulated state, P_{DS}, and are tagged for further investigation of errant activity.

$$P_g \cap GG \to P_{DS}$$

The second form of evidence is scoring GISTIC genes on the basis of contribution to perturbation, high-throughput data agreement, and graph structure placement. Since Fisher's exact test is simply another tool for ORA, it faces the same limitations as previously discussed for enrichment analysis of differentially expressed genes in that the graph structure and placement of these genes is neglected. Thus, the true impact that GISTIC genes have in potentially altering a pathway may be severely restricted to a simple explanation of being found in the pathway. In order to mitigate these issues, the role that GISTIC genes have in altering pathway activity is captured using a gene influence $GINF$ scoring component. This scoring component measures the influence of each gene on the total accumulated perturbation, Acc, of a pathway as captured by the original SPIA method. The $GINF$ score is formally described as

$$GINF(g_j) = \sum_{i=1}^{n} gInf_{ji} \times \frac{1}{pathwayHops_{g_{ji}}} \times HTA(g_j) \tag{3.10}$$

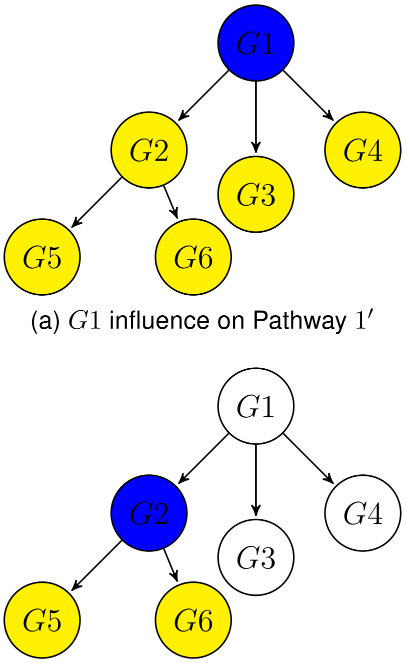
where $gInf_{ji}$ represents the amount of perturbation introduced to the accumulated perturbation at g_i from g_j.

$$gInf_{ji} = \frac{|maxPF(g_j)|}{|Acc'(g_i)|} \tag{3.11}$$

where $maxPF(g_j)$ represents the maximum perturbation factor gene g_j may introduce to the perturbation accumulation measured at any gene g_i in a pathway. Let us assume that in a given pathway, gene g_j is differentially expressed and has only out going edges then finding the maximum perturbation factor, $maxPF(g_J)$, passed to downstream neighbor genes is simply the measured expression of gene g_j divided by the total number of downstream genes $N_{ds}(g_j)$.

$$maxPF(g_j) = \frac{\Delta E(g_j)}{N_{ds}(g_j)} \tag{3.12}$$

Since biological networks and, more specifically, signal transduction networks display scale-free network properties that are sparsely connected [20, 21, 22], the use of the maximum perturbation factor to calculate $gInf$ at any given gene g_i in a pathway ensures simplicity while maintaining a confident standard of accuracy. Let us return to figure 1 to gauge how $GINF$ measures two different genes dependent upon the location with respect to the graph topological structure. The influence that gene $G1$ and gene $G2$ have on pathway $P1'$ can be demonstrated in the following example.



(a) $G1$ influence on Pathway $1'$

(b) $G2$ influence on Pathway $1'$

Figure 3.4: Comparison of influence on pathway $P1'$ between genes

Gene	ΔE	maxPF	GINF
G1	1.5	0.5	3.8
G2	2	1	1.6
G3	0	0	0
G4	0	0	0
G5	0	0	0
G6	0	0	0

Table 3.5: Gene Influence Score for each differentially expressed gene in Pathway $1'$ taking into consideration only $maxPF$

This is a simple yet reasonable approach in calculating the potential influence of gene neighbors on a gene g_i, however, as can be seen in tables 3.2 and 3.3, the maximum perturbation factor of gene g_j continues to trickle down passed its direct neighbor genes. In order to better reflect the potential direct and indirect influence on other genes in a given pathway, $gInf$ is divided by the number of nodes traversed within a pathway from g_i to g_j (hops) as represented by $pathwayHops_{g_{ji}}$. This has two necessary effects on gene influence measurement. The fist is that it rewards genes that are directly connected which is essential in a scale-free topological network. Second, it represents a reasonably assumed degradation of influence as g_j is further separated from g_i.

The last term in the $GINF$ equation ensures agreement between high throughput data being used, which in this case is gene expression measurements and copy number data. This term is referred to as High Throughput Agreement (HTA) and is defined by equation 3.13.

$$HTA(g_j) = 2 \times K^{\frac{S_{gex_j} \times S_{cna_j}}{2}} \tag{3.13}$$

where K is a specified value representative of the confidence of the high-throughput data, S_{gex_j} is a variable related to gene expression data of gene g_j, and S_{cna} is a variable related to the copy number data of gene g_j. In it's simplest form $K = 1$ and ternary values are used for S_{gex} and S_{cna}. Ternary values would represent the state of the gene expression, S_{gex}, at a specified threshold where 1 is up-regulated, -1 is down-regulated, and 0 is normal. Similarly, ternary value would represent the state of the copy number, S_{cna}, at a specified threshold where 1 is gain, -1 is deletion, and 0 is normal. HTA rewards genes that have copy number and gene expression aberration evidence in agreement and slightly penalizes those that have aberration evidence that contradicts each other. Those where a ternary state of 0 is present in the high throughput data are neither penalized nor rewarded. The concept is easily grasped by plotting the ternary states of the high throughput data on a two-dimensional coordinate system as displayed in figure 3.5.

Thus, those points with a positive slope are rewarded, those with a negative slope are slightly penalized, and those with a slope of zero are left alone. Formally, the HTA value is given by taking the distance from a point and raising it to the power of the resultant slope.

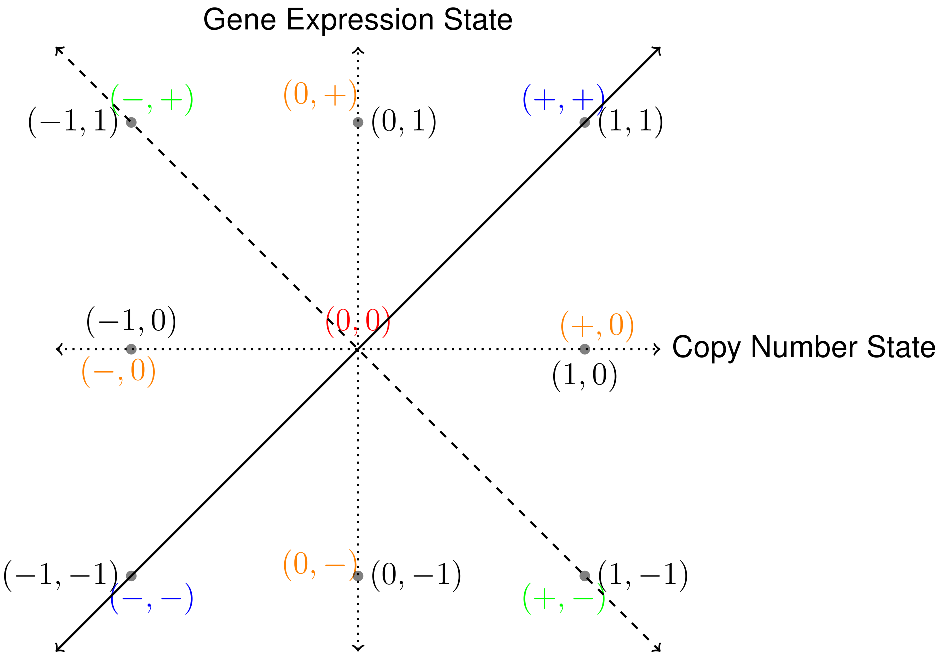


Figure 3.5: States pertaining to high throughput data for a given gene plotted in Cartesian coordinate system

$$HTA(g_j) = (\sqrt{(S_{cna_j} - 0)^2 + (S_{gex_j} - 0)^2})^{\frac{S_{cna_j}-0}{S_{gex_j}-0}}$$

Using a Karnaugh map to display the reward and penalty values for HTA, identifying a simplistic and elegant equation becomes much more intuitive and is used to formulate equation 3.13.

			CNS	
		1	0	-1
	1	1.41	1.00	0.71
GES	0	1.00	1.00	1.00
	-1	0.71	1.00	1.41

Similar to the SPIA method, efficiently calculating the $GINF(g_i)$ score for each gene g_i in a pathway requires the use of matrix data structures.

First, the vector $maxPF$ is computed by simply doing a pairwise division of the ΔE and N_{ds} vectors.

$$maxPF = \Delta E./N_{ds}$$

where N_{ds} is a vector representing the number of downstream genes for each gene in a pathway.

$$N_{ds} = \begin{pmatrix} N_{ds(g_1)} \\ N_{ds(g_2)} \\ \vdots \\ N_{ds(g_n)} \end{pmatrix}$$

The HTA vector is then computed by letting Δ_{Etern} and Δ_{Ctern} represent vectors of S_{gex_j} and S_{cna_j} for ever gene g_j in a pathway using ternary values. Another vector, $2v$, that's size is equal to the number of genes in a pathway containing a constant of 2 is also required.

$$2v = \begin{pmatrix} 2 \\ 2 \\ \vdots \\ 2 \end{pmatrix}$$

$$\begin{aligned} HTA &= \Delta Etern. * \Delta Ctern \\ &= HTA./2v \\ &= 2v.^{HTA} \end{aligned}$$

Element wise multiplication of the vectors $maxPF$ with HTA is then performed.

$$gImp = maxPF. * HTA$$

This vector is expanded where the vector was repeated as a row the number of times needed to satisfy a square matrix.

$$gImp = \begin{pmatrix} gImp_{g_1} & gImp_{g_2} & \cdots & gImp_{g_n} \\ gImp_{g_1} & gImp_{g_2} & \cdots & gImp_{g_n} \\ \vdots & \vdots & \ddots & \vdots \\ gImp_{g_1} & gImp_{g_2} & \cdots & gImp_{g_n} \end{pmatrix}$$

Similarly, the modified Acc' vector is expanded where the vector was repeated as a column the number of times needed to satisfy a square matrix.

$$AccM = \begin{pmatrix} AccM_{g_1} & AccM_{g_1} & \cdots & AccM_{g_1} \\ AccM_{g_2} & AccM_{g_2} & \cdots & AccM_{g_2} \\ \vdots & \vdots & \ddots & \vdots \\ AccM_{g_n} & AccM_{g_n} & \cdots & AccM_{g_n} \end{pmatrix}$$

Element wise division on $gImp$ using $AccM$ is then performed to give matrix C_{GINF}.

$$C_{GINF} = gImp./AccM$$

The matrix C_{GINF} represents what the gene influence for each gene would be if the pathway was represented by a completely connected graph. It is, therefore, necessary to determine the true topological structure of a pathway. The matrix PH consisting of the nodes traversed (hops), ph_{ij}, between genes g_i and g_j provides the necessary topographical information. This information is obtained directly from pathway databases . In order to find the shortest distance (hop count) between each gene, Djikstra's algorithm was used where the default distance was set to ∞ indicating that no path exists from one gene to the next.

$$PH = \begin{pmatrix} ph_{11} & ph_{12} & \cdots & ph_{1j} \\ ph_{21} & ph_{22} & \cdots & ph_{2j} \\ \vdots & \vdots & \ddots & \vdots \\ ph_{n1} & ph_{n2} & \cdots & ph_{nj} \end{pmatrix}$$

Element-wise division is then computed between C_{GINF} and PH.

$$GINFM = C_{GINF}./PH \tag{3.14}$$

The result is a matrix, $GINFM$, where the columns represent the influence that gene g_j has on gene g_i. The sum of each column was computed to obtain the vector $GINF$. This vector represents the score related to the amount of influence that each gene has on the overall accumulation of a pathway.

Once the total $GINF$ score has been found for each gene in a pathway, GISTIC gene scores are compared against normal member gene scores. Using Welch's T-test, a p-value P_{GINF} is obtained that describes the probability of finding the average GISTIC gene influence on a pathway by chance given the average influence of the other member genes. This, ultimately, gives a clearer understanding of the role that GISTIC genes play in pathway activity.

Combining P-values to determine significance of evidence

The p-values generated for each of the evidences used to determine whether a signaling pathway was significantly altered were then combined using the Logit combinational method. Two p-values were specifically used for gene expression analysis while the other two incorporated some form of copy number analysis as shown in table 3.6.

In determining the best method to combine the p-values from the different scoring components, one must fist specify an appropriate null hypothesis then determine the importance of each p-value in accepting or rejecting the null hypothesis.

	Copy Number Evidence	Gene Expression Evidence
Enrichment Analysis	P_{NGG}	P_{NG}
Perturbation Analysis	P_{GINF}	P_{PERT}

Table 3.6: A summary of p-value score generated for each type evidence analyzed

In addition, one must also determine and justify whether each p-value is independent or dependent up on each other. Let us begin with a formal hypothesis as stated here.

H_0 =There are no subset of copy number altered genes that are causing significant aberrant activity in a pathway.

In order to prove the alternative of this hypothesis, one must not only prove a pathway is displaying aberrant activity but also that it is due to copy number altered genes. Therefore each p-value obtained from the scoring method WSPIAGG is important and should play some role in proving the alternative. In addition, the independence of the first two scoring components was justified by Tarca et al. due to the use of the boot strap procedure in computing P_{PERT} and verified through simulation of randomized pathways [9]. This leaves justifying the independence of the final scoring components. Theoretically, it is justified given that copy number data is generated completely separate from gene expression data. Moreover, the method for identifying GISTIC genes do not rely upon any gene expression data. The final scoring component P_{GINF} independence is rooted in the basis that, similar to P_{PERT}, the true independence resides in network topology not gene expression as demonstrated in the previous examples. Therefore, a method for combining independent p-values assigning each a significant level of importance is required

Loughlin et al. compared several meta-analysis methods on different types of data to ascertain which methods scored the best in rejecting a global null hypothesis when combing p-values [23]. Methods were compared changing different parameters of the theoretical data such as number of null hypothesis to combine, evidence distribution, and the strength of evidence in the null hypothesis. For combining three or more p-values, Logit and normal scoring proved to be the most powerful with respect to evidence distribution across all p-values [23]. Upon careful consideration, the Logit method was chosen based on its performance in testing and the ease of implementation.

$$L = \sum_{i=1}^{n} log(\frac{p}{1-p}) \tag{3.15}$$

where n is equal to the number of null hypotheses being tested. The distribution of the Logit function has been shown to be a very close approximation to the normal distribution function with a scaling factor of $d \approx 1.7$ [24].

$$|\Phi(x) - \Psi(dx)| < 0.1 \tag{3.16}$$

where Ψ represents a normal distribution and Φ represents the Logit distribution.

To ensure the most simplicity and strength, the Logit method is what was used to calculate the overall p-value of the null hypothesis.

3.3 Implementation

The theoretical formulation of such a multi-level, intense scoring mechanism must be feasible for efficient implementation and application to real world data. Much of the software used for implementation was proprietary and developed as required. The two main languages that were used in software development were Java and R project. The JRI library allowed for ease of implementation and collaboration between Java and R in implementing the original SPIA method as well as developing novel aspects of the WSPIAGG method. Java was selected for its platform independence and simplicity of use. R project was selected as the main statistical analysis tool for its reliability, power in computation, and robustness in interacting with java.

Netbeans 6.8 was used to develop different graphical user interface for ease of use. JAMA matrix package was heavily relied upon to construct adjacency matrices and assist in the implementation of SPIA in Java. In addition, previously developed software known as EPiCC and ExPattern were used to infer context-specific GRNs from quantized gene expression data.

High Throughput Data

Gene expression and copy number profiles were stored in R objects for ease of access and use where the rows were representative of the corresponding copy number and gene expression values of a given gene, and the columns represented samples that the measurements were taken from. Both gene expression and copy number data were also quantized for use in computing the high throughput agreement value, HTA, and inferring context-specific GRNs.

Identifying Potentially Deregulated Pathways For WSPIAGG Analysis

A precompiled set of previously identified GISTIC genes was required for querying publicly available pathway databases. Only signaling pathways from those databases that offered web services for querying such as Pathway Commons and KEGG were used.

Chapter 4

DATA ANALYSIS AND RESULTS

4.1 Application To Breast Cancer Dataset

In order gauge the effectiveness of the WSPIAGG methodology, breast cancer data containing copy number and gene expression profiles were used. WSPIAGG was compared to SPIA to see how it performed across the different samples.

Tumor Samples Information

Breast cancer tumor samples were obtained from the Southern Sweden Breast Cancer Group tissue bank, Skane University Hospital, Lund, the Helsinki University Central Hospital, and Landspitali University Hospital [25, 5]. The median overall survival follow up time was 8.1 years ranging from 0.24 to 32 years [25, 5]. There were 346 primary tumors and the rest were attributed to local recurrences or lymph node metastases [25, 5].

Gene Expresion Data

The global gene expression profiles of the 359 breast tumor samples consisted of over 10,000 individual probes measuring mRNA using oligonucleotide microarrays (Gene Expression Ominbus, GEO, platform GPL 5345) produced at the SCIBLU Genomic Centre at Lund University [5]. Hybridization, labeling, and image analysis were also all initially conducted at SCIBLU as described in [25]. The expression data was normalized across an additional 218 breast tumor samples and the tumor samples were classified accordingly into six intrinsic molecular subtypes first defined by Hu et al [26]. NCBI entrez id mapping of gene symbols using DAVID [27, 28] and HGNC [29] resulted in identifying ≈ 8200 unique probes as genes in the gene expression profile. The median expression value was taken of those genes that had more than one probe associated with it. The median was decided in order to avoid potential outlying noise. A threshold of ± 1 for normalized gene expression data was used to determine differential expression for WSPIAGG analysis.

Cellular Context Mining On Gene Expression Data

ExPattern software was ran with parameters of $crosstalk = 0.3$ and $interference = 0.1$ in inferring context-specific GRN. This resulted in the identification of 1,977 individual context motifs which were subsequently combined based on gene set membership to form a heavily interconnected context-specific GRN using EPiCC software. Clustering was performed on the context-specific GRN using MCL clustering algorithm with an inflation parameter of 2.0. This value was selected given it's previous performance on cancer data [15]. A total of 94 different clusters of context motifs known as contexts were identified. Samples were then assigned a context or multiple contexts using the sample association score, SAS, with a threshold of < 0.5. Each context was then analyzed for subtype enrichment using the hypergeometric distribution to find the p-value, ps, which represents the likelihood of finding a certain number of samples from a subtype of breast cancer by random chance in a particular context. Contexts that had p-values $ps < 0.001$ were considered enriched with the subtypes of breast cancer represented by the tumor samples. This ps value was used to define the introduced error associated with inter-context hops, ϵ, in equation 3.8.

Copy Number Data

Copy numbers for approximately 32,000 clones were obtained using bacterial artificial chromosomes (BAC) microarrays (GEO platform GPL4723) produced by the SCIBLU Genomics Centre, Lund University, Sweden, for each of the 359 breast cancer tumors. Using R project to map clones to gene regions and DAVID and HGNC for gene symbol to entrez id mapping yielded ≈ 7600 unique genes from the probes. As with gene expression data, hybridization, labeling, image analysis, normalization and break point analysis were conducted on the copy number data at SCIBLU Genomic center as well [25, 5]. Gains and losses were identified by sample adaptive thresholds as described in [25]. Copy number thresholds were set to ± 0.2 in determining copy number deletion and gain for WSPIAGG analysis.

GISTIC Implementation On Copy Number Data

After determining regions of amplification and deletion, the GISTIC algorithm was applied to the tumor samples to identify statistically significant amplification and deletion peaks across the 359 tumor samples. Hierarchical clustering was conducted using Pearson's correlation on significant GISTIC peaks with complete linkage on average scaled $\log_2$ ratio for each peak [5]. Six subtypes labeled Basal-like, Luminal complex, Luminal simple, 17q12, amplifier, and mix, were identified and had significant overlap with the six subtypes identified using Hu et al's method across the 359 tumor samples [5, 26]. Thus, Jonsson et al were able to demonstrate that the genomic landscape as defined by the GISTIC method was capable of grouping tumor samples together that shared similar clinical characteristics. Moreover, well known oncogenes such as MYC, HER2, and MDM2 were located in aberrant regions of importance as well as demonstrating significantly correlated gene expression levels [5].

Basal-like, Luminal complex, 17q12 and Luminal simple were selected for pathway analysis given the clinical characteristics that each shared as well as differed from. From each of the four subtypes a set of genes was identified within the respective GISTIC regions. There were 714 genes found in Basal-like GISTIC regions, 770 genes found in luminal complex GISTIC region, 460 genes found in 17q12 GISTIC regions, and 393 genes found in luminal simple GISTIC regions.

Pathway Repositories

Four publicly available databases were used to obtain pathway information for initial statistical analysis.

Pathway Commons

Pathway Commons database is is a consolidation of other well known pathway databases and is a collaboration between the University of Toronto and the Computational Biology Center at Memorial Sloan-Kettering Cancer Center [30]. Since Pathway Commons consists of pathway information from different databases, the quality of the database is dependent on the quality of the consolidated databases. Some of the more notable databases consolidated within pathway commons are NCI nature, Reactome, and BioGRID. Pathways in Pathway Commons are stored in BIOPax level 2 (BIOlogical PAthway eXchange) format. This database consists of ≈ 5000 verified human genes and ≈ 1200 pathways. As of June 2011, pathways were stored in level 2 format incapable of identifying gene regulatory networks within signaling pathways. In the near future, the repository should transition to BioPAX level 3 to resolve these issues.

Wikipathways

Wikipathways is an open and collaborative repository to create and edit pathways, as the name implies, similar to Wikipedia [31]. It is maintained by BiGCaT Bioinformatics (Maastricht University) and the Conklin Lab at the Gladstone Institutes (University of California, San Francisco). The creation and editing of these pathways is made simple through the use of a graphical editing tool that caters to users with different biological and computer backgrounds. Wikipathways uses GPML (GennMapp Pathway Markup Language) as it's main pathway format. This database consists of ≈ 4300 verified human genes with 369 pathways but given the ease of pathway submission these numbers continue to grow. This, however, is also a hinderance given the lack of standards in pathway development.

KEGG

KEGG (Kyoto Encyclopedia of Genes and Genomes) Pathway database is one of sixteen main databases under the well known and longstanding bioinformatics resource named KEGG [19]. The entire KEGG resource is under the auspices of the Kanehisa Laboratories at Kyoto University and the Human Genome Center of the University of Tokyo. The pathways in the KEGG Pathway database are manually created and stored in KGML (KEGG Graphic Markup Language) format. KEGG Pathways contains ≈ 5000 verified genes with 389 pathways. The ease of access to gene and pathway information makes it ideal for pathway analysis.

Biomodels

Biomodels database is a database that contains curated models that have been peer-reviewed and published [32]. Although this database contains models, the difference between its models and pathways are one of information quantity. Mathematical modeling is used in describing the interactions that occur in a model through SBML (Systems Biology Markup Language) format. Moreover, controlled annotations and related information are also available to the user. Thus, pathways would be considered a single of component of many in a model. Although this database provides much more information on models, and by definition pathways, it contains a relatively small number of pathways compared to the others with only ≈ 600 verified genes and 326 curated pathways. The gene interaction information is also quite cumbersome to extract and not suitable for the type of pathway analysis that is within the scope of this thesis.

Of the four pathway repositories queried, only Wikipathways and KEGG results were selected for further pathways analysis since Pathway Commons and Biomodels were not conducive to WSPIAGG pathway analysis. In the case of pathways commons, it provided a significant amount of pathways but was limited in scope to only protein-protein interaction (PPI) data. This is due to Pathway Commons using BioPAX level 2 and the inherent restrictions within this specific level [33]. Biomodels provided a significant amount of information pertaining to biochemical interactions between molecules but proved to be extremely complex in extracting simple gene interaction information.

The rules defining an edge and its ternary value were based on those in the original SPIA method for KEGG pathways. The task of defining rules proved to be much more complex for Wikipathways given that the original file a pathway is created in is based on graphical syntax. Therefore, an approximation of what constituted an interaction was based on line objects and the shape of their endpoints where arrows indicated activation and t-bars indicated inhibition. Group to group interactions were excluded given that there was a higher probability that group to group indicated a transitioning from one biochemical state to the next as opposed to influence or regulation.

Interaction Type	Edge Value
activation	1
compound	0
binding/association	0
expression	1
inhibition	-1
activation_phosphorylation	1
phosphorylation	0
indirect	0
inhibition_phosphorylation	-1
dephosphorylation_inhibition	-1
dissociation	0
dephosphorylation	0
activation_dephosphorylation	1
state	0
activation_indirect	1
inhibition_ubiquination	-1
ubiquination	0
expression_indirect	1
indirect_inhibition	-1
repression	-1
binding/association_phosphorylation	0
dissociation_phosphorylation	0
indirect_phosphorylation	0

Table 4.1: KEGG interactions definition for edge values

Interaction Type	Edge Value
gene-gene $\rightarrow$	1
gene-gene $\dashv$	-1
group-gene $\rightarrow$	1
group-gene $\dashv$	-1
gene-group $\rightarrow$	1
gene-group $\dashv$	-1
group-group	0

Table 4.2: Wikipathways interactions definition for edge values

Pathways that had an overall calculated p-value ≤ 0.05 were considered significant and were selected for further analysis. Since there is much that is not known of signaling pathway analysis and this type of analysis is still in its infancy, it is best to cast a wider net while maintaining a certain confidence level to allow biologists to make the final determination of significance.

4.2 Results

WSPIAGG was compared against SPIA in two different measurements. First, the average p-value of all pathways was taken across the different samples of each of the cancer subtypes analyzed. Those pathways that had a score for every sample in a subtype were used to compare the two scoring methods. Secondly, the pathway activity detected in the different subtypes were displayed as heatmaps using a p-value threshold of ≤ 0.05 to determine activity status significance. Throughout the different subtypes, a common pattern emerges in which more pathways are found using the original SPIA method but more consistency is found using the WSPIAGG method across the tumor samples for pathway deregulation.

Basal Like Analysis

Comparison Results

Previous research has linked Basal-like subtype to aggressive forms of cancer with a worse prognosis then other subtypes such as luminal simple [5, 26]. Both SPIA and WSPIAGG were able to identify pathways implicated in cancer such as the cell cycle pathway and the focal adhesion pathway. However, WSPIAGG was found to score these pathways much lower than SPIA implicating that the GISTIC genes found in these pathways are complicit in the deregulation.

WSPIAGG was also able to identify more pathways consistently deregulated across the tumor samples then SPIA. This indicates that much more differentiation both genomic and expression wise was picked by WSPIAGG. One explanation is that Basal-like subtype was highly associated with BRCA1 mutated tumors [5]. This

WSPIAGG		**SPIA**	
pathwayName	Ave P-value	PahtwayName	Ave P-value
Non-small cell lung cancer :path:hsa05223	0.0017	Complement and Coagulation Cascades:WP558	0.1006
Prostate cancer :path:hsa05215	0.0192	Complement and coagulation cascades :path:hsa04610	0.1252
Focal Adhesion:WP306	0.0624	ECM-receptor interaction :path:hsa04512	0.1429
Complement and coagulation cascades :path:hsa04610	0.0668	Focal adhesion :path:hsa04510	0.1461
Cell cycle :path:hsa04110	0.1051	Focal Adhesion:WP306	0.1629
DNA damage response (only ATM dependent):WP710	0.1195	Systemic lupus erythematosus :path:hsa05322	0.1862
Focal adhesion :path:hsa04510	0.1252	Pathways in cancer :path:hsa05200	0.1969
Notch signaling pathway :path:hsa04330	0.1486	Cell cycle :path:hsa04110	0.2104
Pathways in cancer :path:hsa05200	0.1567	Small cell lung cancer :path:hsa05222	0.2221

Table 4.3: Comparison between WSPIAGG and SPIA in Basal-like subtype

has implications becasue the BRCA1 gene is responsible for DNA damage repair. Therefore, it is quite plausible to infer that this subtype of cancer will have much more significant chromosomal aberrations associated with it leading to much more consistency being picked up by the WSPIAGG method.

Pathways Of Interest

WSPIAGG identified several pathways that showed consistent activity across tumors samples that have been linked to cancer.

Pathways of particular interest were DNA Damage Response, mTOR signal-

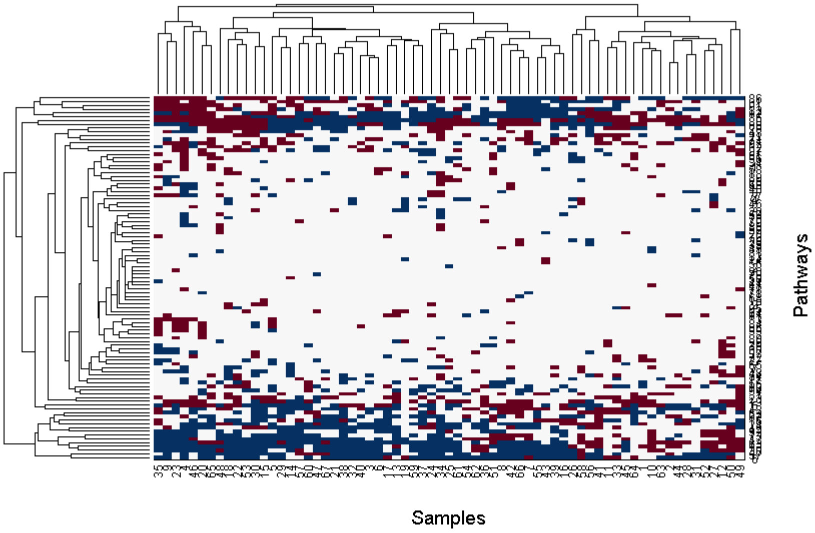


Figure 4.1: Pathway activity as measured by WSPIAGG in Basal-like subtype

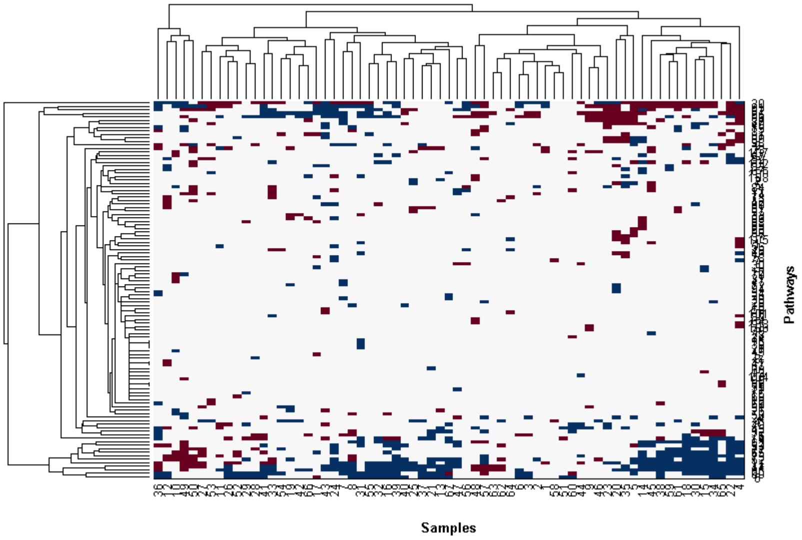


Figure 4.2: Pathway activity as measured by SPIA in Basal-like subtype

ing, and Notch Signaling, all of which have been implicated in cancer and specifically breast cancer in the case of DNA Damage Response [34, 35, 36]. Further investigation into the GISTIC genes in these pathways may be worthwhile.

Notable Pathways	
Pathway Name	Database(s)
G1 To Cell Cycle Control	Wikipathways
Focal Adhesion	KEGG and Wikipathways
Cell Cycle	KEGG and WikiPathways
mTor signaling Pathway	KEGG
Notch Signaling	KEGG
ErbB signaling Pathway	KEGG
DNA Damage Response (only ATM dependent)	Wikipathways
Complement and Coagulation Cascades	KEGG
Antigen Processing and presentation	KEGG
Apoptosis	Wikipathways

Table 4.4: Active pathways in Basal-like subtype

ClusterGroup	Observed	Expected	P-value
0	20	17.74	0.335
1	2	2.79	0.615
2	2	2.12	0.930
3	7	8.21	0.592
4	***9***	***4.61***	***0.0231***

Table 4.5: Logrank test for each clustered group in Basal-like subtype

Overall Survival Analysis

In order to determine whether any of the samples in the Basal-like subtypes that displayed similar pathway deregulation had similar over survival rates, tumor samples were grouped into five main clusters using hierarchical clustering with complete linkage. The logrank test was used to identify the significance between each of the different groups across the entire timeline measured.

Group 4 demonstrated a significant difference between the other groups combined. Using Kaplan Meir's survival probability estimate, group 4 was plotted versus the other groups combined.

The results demonstrated group 4 having a lower overall survival than it's counterparts.

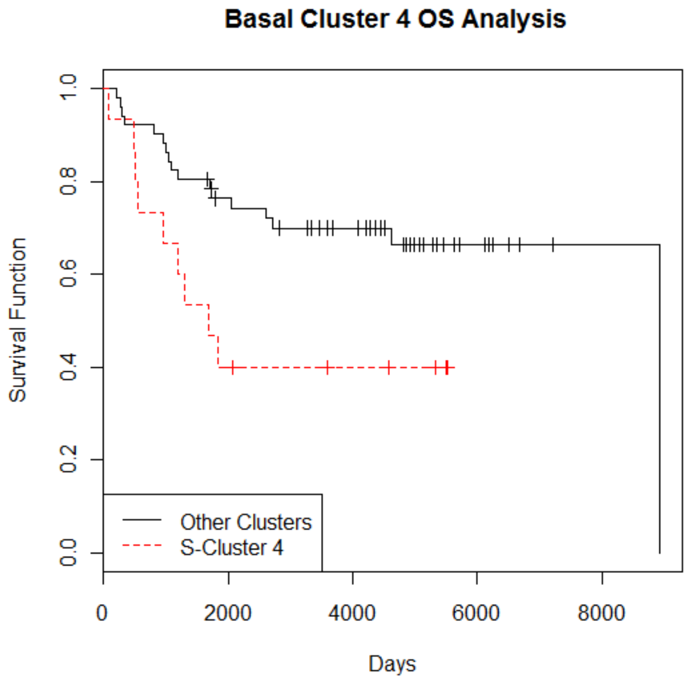


Figure 4.3: Overall survival plots for clustered group 4 in Basal-like subtype

Luminal Complex

Comparison Results

In looking at the average p-values across the samples, WSPIAGG scored worse in most cases than SPIA.

However, it is worth noting two observations. First, WSPIAGG ranked pathways more closely related to cancer higher then SPIA such as Cytokine-cytokine receptor interaction, and Focal adhesion. Second, the pathway activity measured across samples indicates that WSPIAGG found much more pathway deregulation consistently at lower p-values across tumor samples then SPIA.

Pahtways Of Interest

Similar to Basal-like, Luminal complex had notable activity in the Notch Signaling, DNA Damage Response and Focal adhesion which as previously stated have

WSPIAGG		SPIA	
Pathway Name	Ave P-value	Pahtway Name	Ave P-value
Melanoma :path:hsa05218	0.0731	Complement and Coagulation Cascades:WP558	0.0927
Pathways in cancer :path:hsa05200	0.0973	Complement and coagulation cascades :path:hsa04610	0.1024
Cytokine-cytokine receptor interaction :path:hsa04060	0.1994	ECM-receptor interaction :path:hsa04512	0.1692
Focal adhesion :path:hsa04510	0.3102	Systemic lupus erythematosus :path:hsa05322	0.1723
Insulin signaling pathway :path:hsa04910	0.3606	Focal adhesion :path:hsa04510	0.1931
Cell cycle :path:hsa04110	0.3705	Cytokine-cytokine receptor interaction :path:hsa04060	0.2067
Complement and coagulation cascades :path:hsa04610	0.3989	Endochondral Ossification:WP474	0.2169
ECM-receptor interaction :path:hsa04512	0.4277	Focal Adhesion:WP306	0.2176

Table 4.6: Comparison between WSPIAGG and SPIA in Luminal Complex subtype

demonstrated a significant role in cancer [36, 35, 37]. This may indicate that the more aggressive subtypes of cancer deregulate similar signaling pathways. It may be worthwhile to map the GISTIC genes that are shared between Luminal Complex and Basal-like for further analysis. A list of the more notable active pathways associated with cancer for Luminal Complex is provided.

Again, DNA Damage Response was identified as a pathway with consistent activity across tumor samples. Luminal complex subtype has been associated with BRCA2 mutated tumors [5]. Similar to the BRCA1 gene, the BRCA2 gene is responsible for DNA damage repair making it feasible to see significant amount of

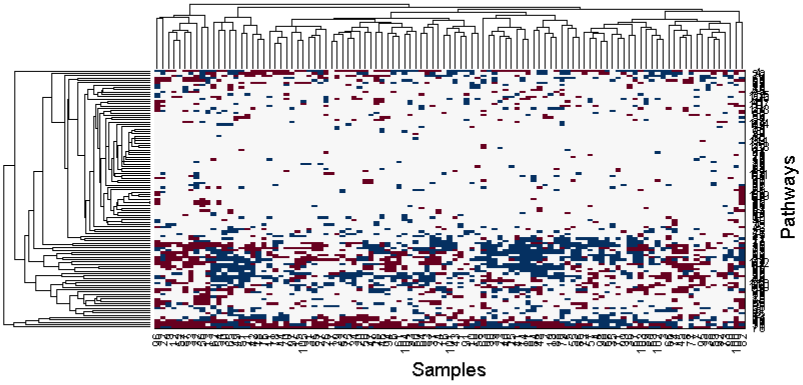


Figure 4.4: Pathway activity as measured by WSPIAGG in Luminal Complex subtype

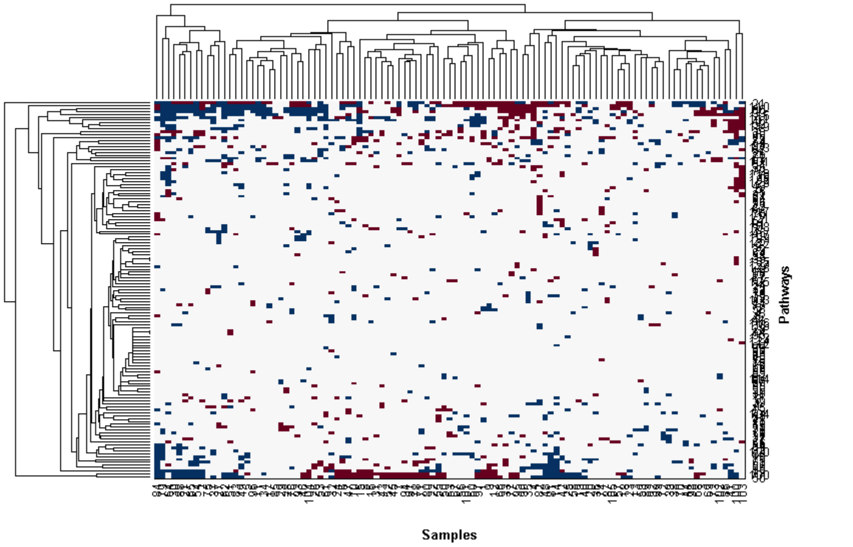


Figure 4.5: Pathway activity as measured by SPIA in Luminal Complex subtype

chromosomal aberrations in these types of tumors with the malfunctioning of the DNA Damage Response pathway.

Overall Survival Analysis

Hierarchical clustering with complete linkage was applied to these samples as well which resulted in five different groups being identified. There was a significant difference between group 4 and the other groups combined with respect to overall survival.

Notable Pathways	
Pathway Name	Database(s)
G1 To Cell Cycle Control	Wikipathways
Focal Adhesion	KEGG and Wikipathways
Notch Signaling	KEGG
DNA Damage Response (only ATM dependent)	Wikipathways
ECM-receptor interaction	KEGG

Table 4.7: Active pathways in Luminal Complex subtype

ClusterGroup	Observed	Expected	P-value
0	34	33.1	0.743
1	8	4.4	0.0573
2	7	3.95	0.111
3	20	24.5	0.237
4	***6***	***13.8***	***0.0107***

Table 4.8: Logrank test for each clustered group in Luminal Complex subtype

However, in this case group 4 appeared to have a higher survival curve then it's counterparts.

Group 1 also demonstrated a slight differentiation compared to the other groups combined displaying a worse overall survival curve but not within an applicable significance range.

17q12

Comparison Results

As within the previous subtypes, SPIA scored better on average across tumor samples then WSPIAGG. However, WSPIAGG as in the other subtypes as well identified much more pathway activity then SPIA.

There wasn't as noticeable the amount of pathway activity as was the case in Basal-like and Luminal complex. However, different pathways still displayed consistent activity across tumor samples.

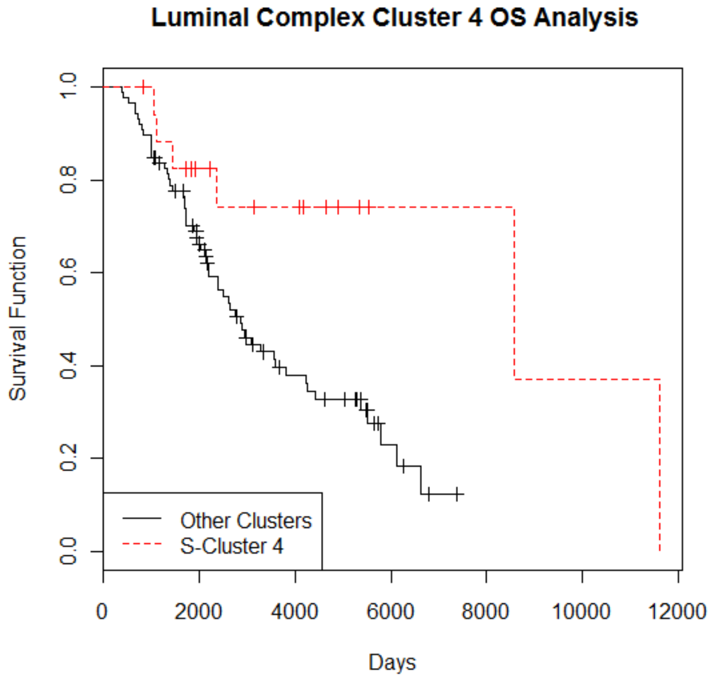


Figure 4.6: Overall survival plots for clustered group 4 in Luminal Complex subtype

Pathways Of Interest

The most notable pathway to be identified using WSPIAGG as active in a significant number of samples was ErbB2 signaling pathway. This is significant because 17q12 shares similar molecular and clinical characteristics to $ErbB2_{+}/Her2_{-}$ subtype of cancer [25].

Insulin signaling pathways have also been implicated in cancer development [38]. Other pathways such as apoptosis have been well established of requiring deregulation in order for tumorigenesis to occur.

Overall Survival Analysis

There were no subgroups in 17q12 subtype that displayed significantly better or worse overall survival differences.

WSPIAGG		SPIA	
Pathway Name	Ave P-value	Pahtway Name	Ave P-value
Small cell lung cancer :path:hsa05222	0.1023	Complement and Coagulation Cascades:WP558	0.0599
Neuroactive ligand-receptor interaction :path:hsa04080	0.2074	Complement and coagulation cascades :path:hsa04610	0.0977
Pathways in cancer :path:hsa05200	0.2595	Cytokine-cytokine receptor interaction :path:hsa04060	0.1560
Insulin signaling pathway :path:hsa04910	0.2679	ECM-receptor interaction :path:hsa04512	0.1611
Cell cycle :path:hsa04110	0.3417	Focal adhesion :path:hsa04510	0.1847
Focal adhesion :path:hsa04510	0.3653	Focal Adhesion:WP306	0.1943
Cytokine-cytokine receptor interaction :path:hsa04060	0.4323	Melanoma :path:hsa05218	0.2789
Prostate cancer :path:hsa05215	0.4494	TGF Beta Signaling Pathway:WP560	0.2979
TGF-beta signaling pathway :path:hsa04350	0.5653	Regulation of actin cytoskeleton :path:hsa04810	0.3056

Table 4.9: Comparison between WSPIAGG and SPIA in 17q12 subtype

Notable Pathways	
Pathway Name	Database(s)
Focal Adhesion	KEGG and Wikipathways
ECM-receptor interaction	KEGG
Insulin signaling pathways	KEGG
ErbB signaling Pathway	KEGG and Wikipathways
Signaling Of Heptocyte Growth Factor Receptor	Wikipathways
Apoptosis	Wikipathways and KEGG
Alpha 6-Beta 4 Integrin signaling pathway	Wikipathways

Table 4.10: Active pathways in 17q12 subtype

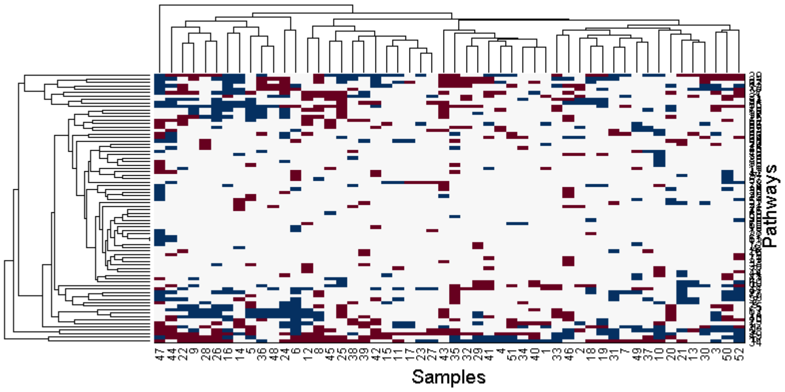


Figure 4.7: Pathway activity as measured by WSPIAGG in 17q12 subtype

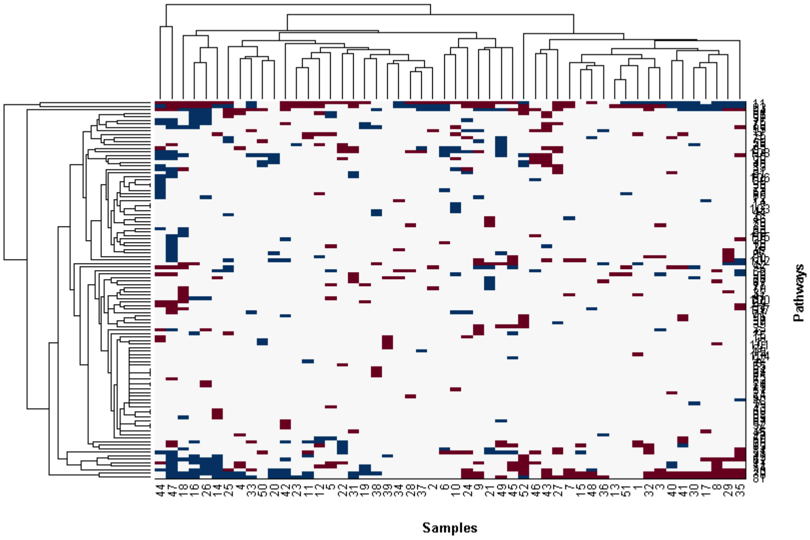


Figure 4.8: Pathway activity as measured by SPIA in 17q12 subtype

Luminal Simple

Comparison Results

Adhering to the same trends, SPIA scored on average better than WSPIAGG across the different samples.

Again, WSPIAGG detected much more consistent activity across samples versus SPIA. Compared to the other subtypes, this subtype did not demonstrate

ClusterGroup	Observed	Expected	P-value
1	4	4.27	0.881
2	6	4.71	0.491
3	3	2.41	0.675
4	5	7.07	0.324

Table 4.11: Logrank test for each clustered group in 17q12 subtype

WSPIAGG		**SPIA**	
Pathway Name	Ave P-value	Pahtway Name	Ave P-value
Cytokine-cytokine receptor interaction :path:hsa04060	0.0824	Cytokine-cytokine receptor interaction :path:hsa04060	0.1846
ECM-receptor interaction :path:hsa04512	0.1209	Systemic lupus erythematosus :path:hsa05322	0.1861
Focal adhesion :path:hsa04510	0.2911	ECM-receptor interaction :path:hsa04512	0.2042
Neuroactive ligand-receptor interaction :path:hsa04080	0.3044	Focal adhesion :path:hsa04510	0.2138
T cell receptor signaling pathway :path:hsa04660	0.5268	Focal Adhesion:WP306	0.2676
Insulin signaling pathway :path:hsa04910	0.5386	TGF Beta Signaling Pathway:WP560	0.2890
Regulation of actin cytoskeleton :path:hsa04810	0.5807	Pathways in cancer :path:hsa05200	0.2968
Cell cycle :path:hsa04110	0.6501	Cell cycle :path:hsa04110	0.3302
Axon guidance :path:hsa04360	0.6951	Cell cycle:WP179	0.3592

Table 4.12: Comparison between WSPIAGG and SPIA in Luminal Simple subtype

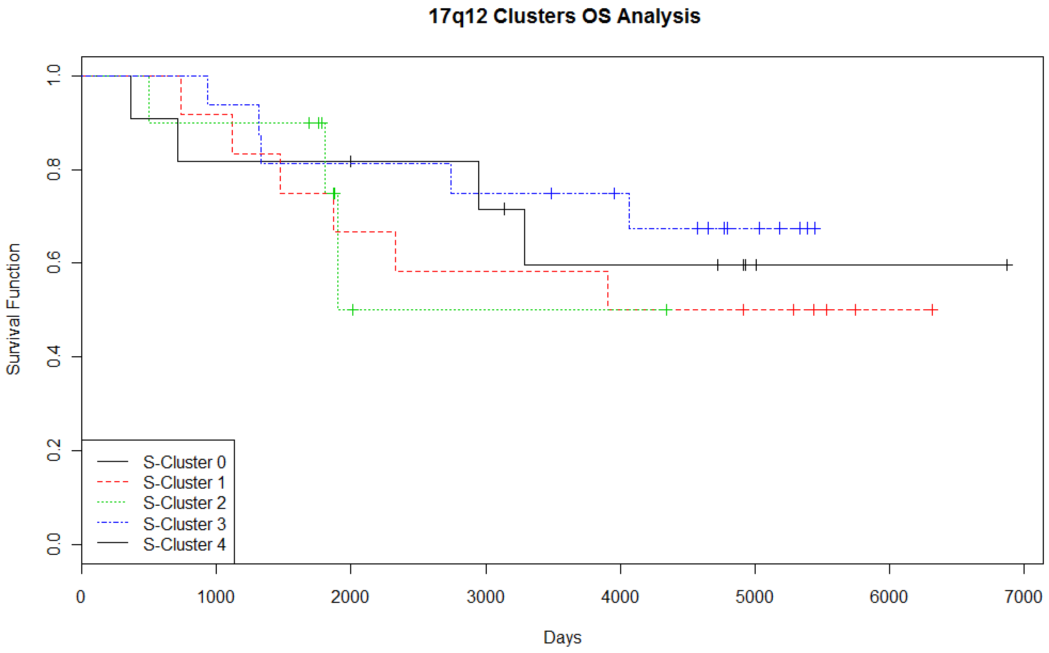


Figure 4.9: Overall survival plots for clustered groups in 17q12 subtype

Notable Pathways	
Pathway Name	Database(s)
Focal Adhesion	KEGG and Wikipathways
ECM-receptor interaction	KEGG
Cytokine-Cytokine Receptor Interaction	Wikipathways
Neuroactie ligand-receptor Interaction	KEGG

Table 4.13: Active pathways in Luminal Simple subtype

large amounts of pathway activity using WSPIAGG method. This may correspond to research evidence suggesting a less aggressive form of breast cancer compared to the other subtypes [26, 25].

Pathways Of Interests

Corresponding to pathway activity, not many pathways were identified as notable.

Cytokine-Cytokine Receptor Interaction has been implicated in cancer [39] and apoptosis is a well known pathways that must be deregulated for tumorigenesis to occur.

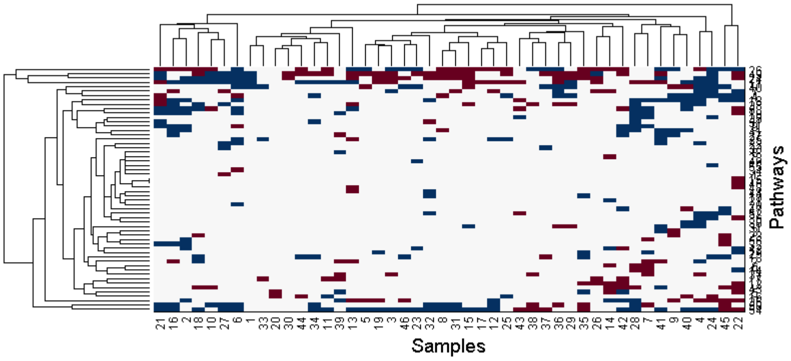


Figure 4.10: Pathway activity as measured by WSPIAGG in Luminal Simple subtype

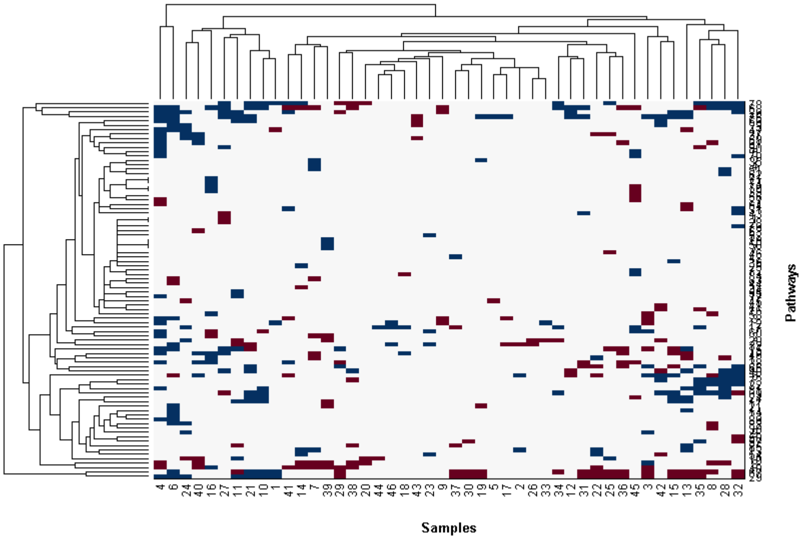


Figure 4.11: Pathway activity as measured by SPIA in Luminal Simple subtype

Overall Survival Analysis

Luminal Simple did not have any subgroups that faired better or worse with respect to overall analysis.

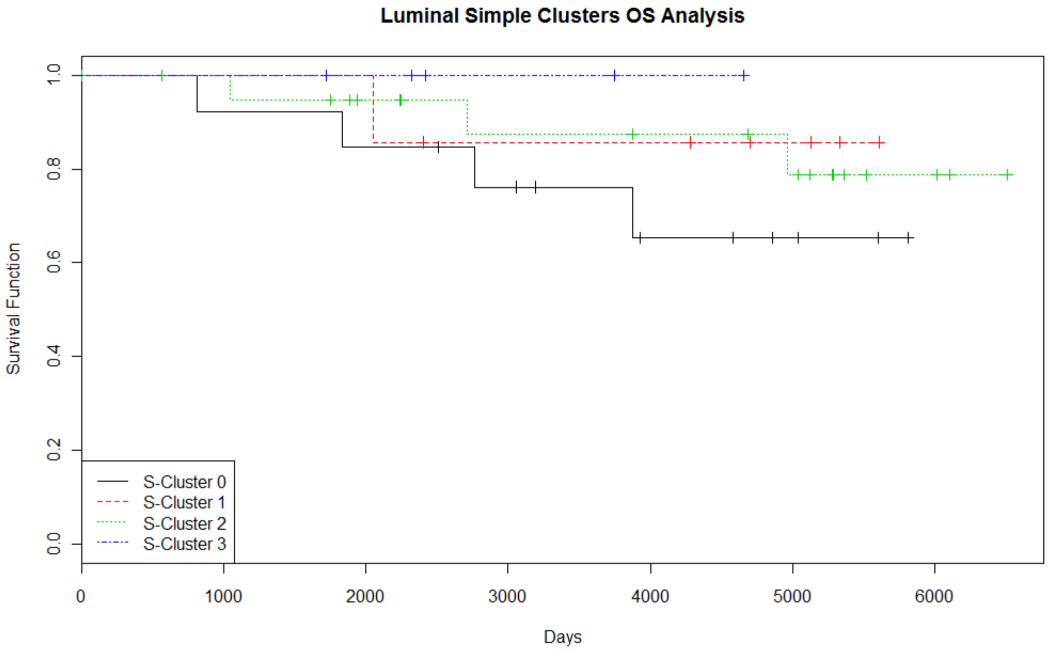


Figure 4.12: Overall survival plots for clustered groups in Luminal Simple subtype

ClusterGroup	Observed	Expected	P-value
0	4	5.68	0.188
1	1	1.41	0.702
2	3	3.66	0.638
3	0	0.614	0.324

Table 4.14: Logrank test for each clustered group in Luminal Simple subtype

Chapter 5

DISCUSSION

The pathways identified as notable in the different subtypes are a who's-who of pathways implicated in cancer. Since there's an argument to be made that every signaling pathway may be implicated, WSPIAGG not only identifies these pathways but points to a solid source, in the GISTIC genes, of what could be causing deregulation. As a point of interest, the same two pathways were consistently identified in the different subtypes of breast cancer. Focal Adhesion and ECM-receptor interaction were found in all four of the subtypes. Further research of these pathways may be useful in breast cancer.

The use of WSPIAGG was of mixed results compared to SPIA. On one hand, SPIA had lower p-values for pathway activity then WSPIAGG on average across the majority of subtypes. On the other hand, WSPIAGG was capable of identifying more pathways as being deregulated more consistently across tumor samples in each of the four subtypes of breast cancer analyzed. This may be explained by the fact that WSPIAGG is reliant upon GISTIC gene activity to determine pathway activity. If GISTIC gene activity is low or if the number of GISTIC genes found in a pathway is not significant then it can severely effect the overall score given to a pathway. Thus, although more pathways were found across the samples to have a lower p-value for SPIA, there could have been a certain number of pathways that did not have significant GISTIC activity or presence that skewed the p-value for WSPIAGG. Given that significant mutations and differential expression are high in the Basal-like subtype, consistent with clinical research, this may explain why the Basal-like subtype was the only subtype to have pathways p-value scores lower on average using WSPIAGG.

Nevertheless, WSPIAGG was capable of demonstrating significant results in identifying pathways associated with cancer for each of the different subtypes analyzed. In particular, identifying ErbB2 signaling pathways in the 17q12 analysis. It also demonstrated the ability to identify the same pathways as SPIA but with much more consistency of lower p-values across tumor samples. Ultimately, the overall objective was to introduce copy number data into signaling pathway analysis and obtain just as good, if not better, results. Given the results previously introduced, copy number data should be taken into consideration when determining pathway deregulation.

Future Research

The current research focused on using KEGG and Wikipathways databases limiting the scope to the information they provided. As Pathway Commons migrates to BioPAX level 3, it will be possible to incorporate a much greater number of signaling pathways to analyze. In addition, as more databases move to a centralized method of storing and representing signaling pathway data, the ease of implementing new databases in proprietary software developed will allow for better and more efficient analysis of signaling pathways.

In addition, proteomics continues to expand the amount of new information with respect to signal transduction and signaling pathways. The inclusion of the data provided by the proteomics field will be of vital importance, especially in signaling pathway analysis. Signaling pathways rely upon a number of different proteins from ligands to enzymes, to ensures proper intercellular and intracellular communication. Incorporating this data into signaling pathway analysis will undoubtedly assist in describing pathway regulation and deregulation.

Finally, as copy number data availability continues to increase for cancer datasets, it will be possible to apply this method to other types of cancer as well. This may prove useful in better understanding the role that gene aberrations play in deregulated pathways across different types of cancer with a sincere hope that it leads to better, more efficient cancer therapies.

REFERENCES

[1] R. Beroukhima et al., “Assessing the significance of chromosomal aberrations in cancer: Methodology and application to glioma,” *PNAS*, vol. 105, no. 50, pp. 20 007–20 012, December 2007.

[2] D. Hanahan and R. A. Weinberg, “The hallmarks of cancer,” *Cell*, vol. 100, pp. 57–70, January 2000.

[3] B. Alberts, A. Johnson, J. Lewis, M. Raff, K. Roberts, and P. Walters, *Molecular Biology Of The Cell*, 4th ed. Garland Science, 2002.

[4] D. Hanahan and R. A. Weinberg, “Hallmarks of cancer: The next generation,” *Cell*, vol. 144, no. 5, pp. 646–674, March 2011.

[5] G. Jönsson et al., “Genomic subtypes of breast cancer identified by array-comparative genomic hybridization display distinct molecular and clinical characteristics,” *Breast Cancer Research*, vol. 12, no. R42, 2010.

[6] P. Khatri et al., “Profiling gene expression using onto-express,” *Genomics*, vol. 79, pp. 266–270., 2002.

[7] T. Breslin, M. Krogh, C. Peterson, and C. Troein, “Signal transduction pathway profiling of individual tumor samples,” *BMC Bioinformatics*, 2005.

[8] Y. Liu and M. Ringnér, “Revealing signaling pathway deregulation by using gene expression signatures and regulatory motif analysis,” *Genome Biology*, vol. 8, no. R77, 2007.

[9] A. L. Tarca, S. Draghici, P. Khatri, S. S. Hassan, P. Mittal, J.-s. Kim, C. J. Kim, J. P. Kusanovic, and R. Romero, “A novel signaling pathway impact analysis,” *Bioinformatics*, vol. 25, no. 1, pp. 75–82, 2009.

[10] E. R. Dougherty, M. Brun, J. M. Trent, and M. L. Bittner, “Conditioning-based modeling of contextual genomic regulation,” *IEEE/ACM TRANSACTIONS ON COMPUTATIONAL BIOLOGY AND BIOINFORMATICS*, vol. 6, no. 2, 2009.

[11] S. Kim, I. Sen, and M. Bittner, “Mining molecular contexts of cancer via in-silico conditioning,” *Computational Systems Bioinformatics Conference*, vol. 6, pp. 169–179, 2007.

[12] G. Palla, I. Derenyi, I. Farkas, and T. Vicsek, “Uncovering the overlapping community structure of complex networks in nature and society,” *Nature*, vol. 435, no. 7043, pp. 814–818, 06 2005. [Online]. Available: http://dx.doi.org/10.1038/nature03607

[13] M. Girvan and M. E. J. Newman, “Community structure in social and biological networks,” *PNAS*, vol. 99, no. 12, pp. 7821–7826, June 2002.

[14] I. Sen, M. P. Verdicchio, S. Jung, R. Trevino, M. Bittner, and S. Kim, “Context-specific gene regulations in cancer gene expression data,” *Pacific Symposium On Biocomputing*, pp. 75–86, August 2009.

[15] A. Ramesh, R. Trevino, D. D. Von Hoff, and S. Kim, “Clustering context-specific gene regulatory networks,” *Pacific Symposium On Biocomputing*, vol. 15, pp. 444–455, 2010.

[16] P. N. Tan and M. S. Vipin Kumar, *Introdcution To Data Mining*. Pearson Education, Inc, 2006.

[17] S. van Dongen, “A cluster algorithm for graphs,” National Research Institute for Mathematics and Computer Science, Tech. Rep., 2000.

[18] J. Shi and J. Malik, “Normalized cuts and image segmentation,” *IEEE TRANSACTIONS ON PATTERN ANALYSIS AND MACHINE INTELLIGENCE*, vol. 22, no. 8, August 200.

[19] M. Kanehisa, “Toward pathway engineering: a new database of genetic and molecular pathways,” *Science And Technology Japan*, no. 59, pp. 34–38, 1996.

[20] M. Arita, “Scale-freeness and biological networks,” *JB Minireview-Bioinformatics And Systems Biology*, vol. 138, pp. 1–4, May 2005.

[21] G. A. Pavlopoulos et al., “Using graph theory to analyze biological networks,” *BioData Mining*, vol. 2, p. 27, 2011.

[22] R. Albert, “Scale-free networks in cell biology,” *Journal of Cell Science*, vol. 118, pp. 4947–4957, 2005.

[23] T. M. Loughin, “A systematic comparison of methods for combining p-values from independent tests,” *Computational Statistics And Data Analysis*, vol. 47, pp. 467–485, November 2003.

[24] G. Camilli, “Teacher’s corner: Origin of the scaling constant d = 1.7 in item response theory,” *Journal of Educational and Behavioral Statistics*, vol. 19, no. 293, 1994.

[25] G. Jonsson et al., “Highresolution genomic profiles of breast cancer cell lines assessed by tiling bac array comparative genomic hybridization.” *Gene Chromosomes Cancer Chromosomes Cancer*, vol. 46, pp. 543–558, 2007.

[26] Z. Hu et al., “The molecular portraits of breast tumors are conserved across microarray platforms,” *BMC Genomics*, vol. 7, no. 96, April 2006.

[27] D. Huang, B. Sherman, and R. Lempicki, “Bioinformatics enrichment tools: paths toward the comprehensive functional analysis of large gene lists,” *Nucleic Acids Res.*, vol. 37, no. 1, pp. 1–13, 2009.

[28] D. Huang, B. Sherman, and R. Lempicki, “Systematic and integrative analysis of large gene lists using david bioinformatics resources.” *Nature Protoc*, vol. 4, no. 1, pp. 44–57, 2009.

[29] Hgnc database, hugo gene nomenclature committee (hgnc), embl outstation. [Online]. Available: www.genenames.org.

[30] A. Cerami et al., “Pathway commons, a web resource for biological pathway data. nucl. acids res.” *Oxfords Journals*, 2010.

[31] A. Pico, T. Kelder, M. van Iersel, K. Hanspers, B. Conklin, and C. Evelo, “Wikipathways: Pathway editing for the people.” *PLoS Biol*, vol. 6, no. 7, 2008.

[32] D. Li et al., “Biomodels database: An enhanced, curated and annotated resource for published quantitative kinetic models,” *BMC Syst Biol*, vol. 4, no. 92, 2010.

[33] E. Demir et al., “Biological pathway exchange (biopax),” *Nature Biotechnology 28*, vol. 10, no. 1038, pp. 935–942, 2010.

[34] E. Petroulakis, Y. Mamane, O. Le Bacquer, D. Shahbazian, and N. Sonenberg, “mtor signaling: implications for cancer and anticancer therapy,” *British Journal of Cancer*, 2005.

[35] V. Dapic, M. A. Carvalho, and A. N. A. Monteiro, “Breast cancer susceptibility and the dna damage response,” *Cancer Control: Journal of the Moffitt Cancer Center*, vol. 12, no. 2, 2005.

[36] B. J. Nickoloff, B. A. Osborne, and L. Miele, “Notch signaling as a therapeutic target in cancer: a new approach to the development of cell fate modifying agents,” *Oncogene*, vol. 22, no. 42, pp. 6598–6608, print 0000. [Online]. Available: http://dx.doi.org/10.1038/sj.onc.1206758

[37] H. Sawai, Y. Okada, H. Funahashi, Y. Matsuo, H. Takahashi, H. Takeyama, and T. Manabe, “Activation of focal adhesion kinase enhances the adhesion and invasion of pancreatic cancer cells via extracellular signal-regulated kinase-1/2 signaling pathway activation,” *Molecular Cancer*, vol. 4, no. 1, p. 37, 2005. [Online]. Available: http://www.molecular-cancer.com/content/4/1/37

[38] V. N. Anisimov, “Insulin/igf-1 signaling pathway driving aging and cancer as a target for pharmacological intervention,” *Experimental Gerontology*, vol. 38, no. 10, pp. 1041 – 1049, 2003. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0531556503001694

[39] L. V. Rhodes et al., “Cytokine receptor cxcr4 mediates estrogen-independent tumorigenesis, metastasis, and resistance to endocrine therapy in human breast cancer,” *Cancer Research*, vol. 71, pp. 603–613., 2011.