Investigating Tumour Evolution Through Graph Theoretical Analysis of Gene Regulatory Networks

Alex Upton

A thesis submitted to

The University of Birmingham

for the degree of

DOCTOR OF PHILOSOPHY

School of Electronic, Electrical and
Computer Engineering
The University of Birmingham
16th May 2014

Abstract

The main aim of this research study was to develop methods to aid biologists and clinicians investigate the progression and evolution of tumours, through the analysis of microarray data. This thesis concentrates on the inference and analysis of Gene Regulatory Networks that represent different evolutionary and clinical stages of cancer cell line microarray data. Three main areas of work were carried out.

The first was the development and implementation of a network inference method specifically designed to infer Gene Regulatory Networks at differently defined classes from a single microarray dataset. Furthermore, this method was shown to be transferable across the different microarray datasets used in this work; and has been deployed on a local host version of a web-based bioinformatics tools server allowing easy access and use.

The second was the investigation of appropriate graph theory metrics to quantitatively analyse the different defined stages of disease. Genes identified by the various metrics were scored for the particular disease of interest using a text-mining tool, allowing the graph theory metrics to be ranked against each other for the various GRNs. This enabled graph theory metrics to be scored for different categories of Gene Regulatory Networks inferred from different cancer microarray datasets, aiding in the identification of appropriate graph theory metrics to apply to Gene Regulatory Networks inferred for different stages of disease.

The third was the comparison of Gene Regulatory Networks inferred for different disease stages across datasets for the same disease, neuroblastoma, from two different studies. Gene Regulatory Networks for common sample clinical disease stages across two different neuroblastoma microarray datasets were inferred using the novel network inference method,

with a high number of common genes identified in the top 100 ranking genes in one of the disease stages, 4M, across both neuroblastoma microarray datasets, highlighting the transferability of the network inference method.

It has been seen in this work that the application and analysis of Gene Regulatory Networks inferred using a method specifically designed to infer multiple GRNs from a single microarray dataset has specifically identified genes involved in different disease or evolutionary stages of disease, and thereby has the potential to aid in the investigation of the progression and evolution of tumours.

In loving memory of my Grandfather, Brian George Upton

Acknowledgements

Firstly, I would like to thank my main supervisor Professor Theodoros Arvanitis for his help over the course of this research and in his assistance in building a cohesive thesis. Thanks are also due to my other supervisors Dr Carmel McConville and Dr Andrew Peet for their valuable input relating to the medical side of things.

Thanks to all the friends and colleagues at the university who have provided support and encouragement over the course of the PhD. I am also grateful to the Engineering and Physical Sciences Research Council for funding the first three years of my work.

I would also like to thank Universitas 21 for providing me with a travel award to spend a month researching at the University of Auckland. Thanks are also due to all those at Auckland who made me feel so welcome, in particular Associate Professor Cristin Print, and Dr Daniel Hurley.

And finally, last but not least, thank you to my friends and family, who have been incredibly supportive and understanding. A special thank you is due to my girlfriend Carla Aloè, for her patience and enthusiasm, especially during the last few months.

Contents

Appendices

List of Figures

List of Tables

List of Common Abbreviations

ARACNE	Algorithm for the Reconstruction of Accurate Cellular Networks
CGPrio	Cancer Gene Priorization
DAVID	Database for Annotation, Visualization and Integrated Discovery
DNA	Deoxyribonucleic acid
dreAm	Dialogue for Reverse Engineering Assessments and Methods
FDR	False Discovery Rate
GATHER	Gene Annotation Tool to Help Explain Relationships
GRN	Gene Regulatory Network
IntOGen	Integrative OncoGenomics
KEGG	Kyoto Encyclopedia of Genes and Genomes
mRNA	Messenger RNA
PPI	Protein-Protein Interaction
R	R programming language
RNA	Ribonucleic acid
WGCNA	Weighted Correlation Network Analysis

Chapter 1

Introduction

1.1 Background

In recent years, there has been an explosion in the availability of high throughput biological data. As a result of this, there has been a shift away from the traditional molecular biology divide-and-conquer reductionist approach, that complex problems are solvable by dividing them into smaller and simpler units [1]. Whilst this reductionist approach has been responsible for tremendous success in the field of medicine, and a great deal of our current understanding of biology, there are limits to the usefulness of this approach. In living organisms, a vast number of biological processes involve the interaction of different biological components, something that cannot be understood using the reductionist approach. As such, an alternative systems level perspective is required in order to complement it [2]. This alternative explanation, the systems perspective, has arisen from systems biology. The aim of systems biology is to understand biological systems at a systems level, by building models of biological systems from information about their components [3].

The interaction of biological components in an organism occurs through various pathways, such as Gene Regulatory Networks (GRNs) [4]. GRNs involve the interaction of different genes in an organism, providing a mechanism to control protein levels in cells. GRNs can be inferred from high throughput microarray data using a number of computational tools, and will be referred to as network inference methods for the rest of this work [5]. One particular systems biology approach has been to infer GRNs from cancer cell line microarray data, such as the studies by Bonnet et al [6], Jeong et al [7], and Horvath et al [8], and use graph theory metrics such as the number of connections a gene has to identify particular genes of interest in these networks. These studies have grouped all the samples in the microarray dataset together, and have inferred a single GRN.

In this thesis, the focus is on distinguishing the samples in cancer cell line microarray datasets into different evolutionary or clinical classes, and subsequently inferring and analysing GRNs for each of these distinct classes. Important information about the different evolutionary and clinical stages can be lost by grouping all the samples together, such as survival time, and inferring a single GRN. Furthermore, by inferring GRNs for different evolutionary and clinical stages, a picture of how the genetic interactions in a disease evolve can be formed. Current network inference methods have been predominantly developed with the goal of inferring a single GRN from microarray datasets, and are therefore not designed for multi-class GRN inference from a single microarray dataset. A method specifically developed for this aim is presented, with application to a number of different cancer microarray datasets. Finally, this method is coded into a web-based platform, allowing simple and unrestricted use by biologists and clinicians, removing the need to learn specialist programming languages that have been a barrier to further adoption of GRN inference and analysis tools.

1.2 Challenges

Three major challenges are identified in the inference and analysis of GRNs from cancer cell line microarray data. These are listed below; from which the objectives of the thesis, set out in section 1.3, are derived.

1. Whilst there is a large amount of existing tumour microarray data from a vast number of cancer cell lines, issues exist in inferring relevant evolutionary information from tumour microarray datasets using GRN inference and analysis methods. Most

microarray datasets include additional information, such as disease class or survival time, by which to categorise the samples. However, current methods are focused on the inference of a single GRN from a microarray dataset. This potentially leads to important evolutionary or clinical stage information being lost as a result; a gene involved in the regulation of a number of other genes at one particular clinical stage, may not be involved in the regulation of the same genes at a different clinical stage. By capturing gene interactions at a number of different stages, a more complete picture of the gene interactions across the evolution of the disease can be formed. As existing GRN inference methods have been developed to infer a single GRN from a microarray dataset, they are not suited to infer multiple GRNs, and do not exhibit the same level of performance across different microarray datasets; performance tends to be heavily biased towards the dataset used to develop the method. GRN inference and analysis methods also require specialist programming skills, that whilst not posing an issue for researchers accustomed to working with computer programming languages, does cause a problem for biologists and clinicians without these specialist skills that wish to make use of these tools. Due to these specialist programming skills required, current methods for GRN inference and analysis are not being fully exploited [9]. It is preferable for biologists and clinicians to be able to directly utilise network inference and analysis tools as they possess the specialist biological and clinical knowledge to be able to interpret the results, rather than computer scientists and engineers.

2. Due to the current focus on inferring a single GRN from a microarray dataset, appropriate graph theory metrics that can be used to quantitatively analyse different evolutionary categories of GRNs have not been defined. Whilst certain graph theory metrics, such as degree centrality and betweenness centrality, have been applied to

single GRNs inferred from a microarray dataset to identify genes of interest, the application of appropriate graph theory metrics to identify genes of interest from different GRNS from a single microarray dataset has yet to be studied. As disease evolves, the ability of the same metric to identify important genes for the disease may alter. An objective overview of which graph theory metrics perform better in the different GRNs inferred for the different stages of disease is therefore required in order to aid researchers interested in inferring and analysing GRNs for different stages of disease. It is likely that a metric that performs well and is able to identify genes already associated with a certain disease, will be a good choice of metric to identify new genes of interest for that same disease.

3. As the focus to date has been on inferring a single GRN from a microarray dataset, the comparison of GRNs inferred for common disease stages from different microarray datasets for the same disease has not been investigated. Comparing the same disease stage GRN across different microarray datasets may be a better indicator to highlight genes that are involved in that disease stage, rather than simply considering one tumour microarray dataset in isolation. If the same genes are identified across multiple datasets, there is likely to be greater confidence in the results, than if genes are identified from one dataset. Therefore, comparing and analysing GRNs for the same disease stage across different microarray datasets could act as an important guide to genes that are important to that disease stage.

1.3 Research Aims and Objectives

The aim of this thesis is to aid biologists and clinicians investigate the progression and evolution of tumours, through the inference and analysis of GRNs that represent different evolutionary and clinical stages of cancer cell line microarray data. In order to do this, the following three objectives have been implemented.

1. Implement a transferable network inference method specifically designed for the inference of different GRNs from a single microarray dataset. Additionally, this method should be easily accessible for biologists and clinicians to carry out gene regulatory network inference and analysis themselves, without needing specialist programming skills and expertise.

2. Investigate appropriate graph theory metrics to quantitatively analyse the GRNs inferred for the different defined stages of disease.

3. Compare GRNs inferred for different stages of disease across microarray datasets from different studies for the same disease.

1.4 Contributions

The thesis offers a number of novel contributions to the analysis of cancer cell line microarray data using GRNs. The first is the implementation of a network inference method specifically designed to infer GRNs at differently defined classes from a single microarray dataset. This method is transferable across different microarray datasets; in this thesis it has been applied to four different microarray datasets, and the application of graph theory metrics to the GRNs inferred has been able to identify unique genes of interest for the different categories of GRN within each microarray dataset.

Furthermore, all network inference methods used in this thesis, including the novel network inference method developed, have been implemented on a local host version of the Galaxy [10] bioinformatics tools web-based server. This implementation allows users to easily access and use GRN inference and analysis tools that previously required specialist computer programming language skills.

The second contribution is the investigation of appropriate graph theory metrics to quantitatively analyse the different defined stages of disease. Genes that various graph theory metrics have identified have been scored using Génie [11], a text mining tool that scores genes for particular biological topics, allowing the graph theory metrics to be ranked against each other for the various GRNs. This has allowed graph theory metrics to be scored for different categories of GRN inferred from different cancer microarray datasets, aiding in the identification of appropriate graph theory metrics to apply to GRNs inferred for different stages of disease.

Building on the second contribution of the work, the third contribution is the comparison of GRNs inferred for different stages of disease across datasets for the same disease from different studies. GRNs for common sample clinical disease stages across two different neuroblastoma microarray datasets have been inferred using the novel network inference method, with a high number of common genes identified in the top 100 ranking genes in disease stage 4M across both neuroblastoma microarray datasets, highlighting the transferability of the network inference method.

1.5 Thesis Structure

The thesis is set out as follows. Chapter 2 gives an overview of the research area. This begins with an introduction to graph theory; in this section various network metrics are discussed, and existing examples of the use of graph theory to model data are shown. Biological networks are also introduced.

Chapter 3 provides a summary of GRNs, and different network inference methods that can be used to construct GRNs from microarray data are explained. Finally, an overview of the proposed approach for this work is presented.

Chapter 4 details the application of an existing network inference method, WGCNA, to the problem of inferring different classes of GRNs from a single microarray dataset. Five different classes of GRNs are inferred from a glioblastoma dataset, using survival time as the class discriminator. Some of the issues that can result from applying existing techniques for inferring GRNs from microarray data are evident in the results, such as the identification of common genes of interest across markedly different survival categories. This arises primarily due to the network inference technique being designed for inferring a single GRN from a microarray dataset, and not distinguishing between the different survival categories.

Chapter 5 details the application of a different network inference method; a novel z-score based approach, to the glioblastoma dataset introduced in chapter 4. The purpose of this chapter is to highlight the application of a network inference method specifically developed for inferring multiple GRNs from a single microarray dataset that addresses the first objective.

Chapter 6 details the application of the novel z-score based approach to two neuroblastoma datasets from different studies. The work in this chapter addresses the second and third objectives, graph theory metrics are scored based on the genes they identify and GRNs that represent different disease stages from the two neuroblastoma studies are compared.

Chapter 7 details the application of a further refined version of the novel z-score based approach to a proprietary retinoblastoma microarray dataset. Samples from normal retinal data are contained in this dataset, and as such, a sample reference network for normal retinal data is also constructed to compare to the retinoblastoma networks.

Chapter 8 details the implementation of GRN inference and analysis tools on a web-based server, including the WGCNA and novel z-score method presented in this work. This further addresses the first objective, by providing an easy-to-use and easily accessible means for biologists and clinicians to infer and analyse GRNs from microarray data.

Chapter 9 provides a summary of the work in the thesis. This includes a discussion on how well the original objectives have been addressed, limitations of the work, and speculates on how the work presented in this thesis could be developed in the future.

There are a number of tables and figures referred to in this work. Where possible, these have been placed in the appendix to improve the readability.

1.6 Summary

The aim of this thesis is to aid biologists and clinicians in the understanding of the evolution and progression of tumours. This can be achieved through the application of the network inference method initially presented in chapter 5 that infers different GRNs from a single microarray dataset, allowing GRNs that represent either different evolutionary or clinical stages to be inferred from a single microarray dataset. Appropriate graph theory metrics for the analysis of these GRNs is then investigated in Chapter 6, along with the comparison of GRNs representing different clinical disease stages for the same disease, neuroblastoma, from two different microarray studies. Chapter 7 presents refined version of the network inference method, and chapter 8 then details the implementation of a number of network inference and analysis tools, including all those used in this work, on a web-based server allowing clinicians and biologists easy-to-use and easily accessible means to investigate microarray data using GRN inference and analysis tools.

Chapter 2

Systems Biology, Biological Networks, and Graph Theory

In this chapter, a brief introduction to systems biology is presented. Following on from this, the concept of biological networks is introduced. GRNs, introduced briefly in the previous chapter, are further expanded on, along with other types of biological network. Relevant areas of graph theory are then finally introduced that can be used to model and analyse these biological networks, with a particular focus on graph theory concepts that are applicable to GRNs.

2.1 Systems Biology

In recent years, there has been a realisation that there are limits to be usefulness of the reductionist approach traditionally advocated in biology. This divide-and-conquer approach, that complex problems are solvable by dividing them into smaller and simpler units, studied the properties of the cell on an individual molecular basis. This traditional approach of reductionism has been responsible for successfully identifying most of the cellular interactions and components, and thus heralded a huge number of key breakthroughs in the history of biology, such as the existence of genes and the discovery of DNA. However, the reductionist approach does not offer any methods to understand how system properties emerge [1], and therefore an alternative approach is required in order to complement it [2].

This has caused a shift in the approach used in recent years, with the realisation that emergent properties can be discovered and better understood by taking a systematic view of biological processes, through observation, using quantitative measures, of multiple components and also data integration with mathematical models. A key reason for this is that within a cell, biological functions rarely arise due to individual molecules; instead they take place because of interactions between many different molecules. Therefore, biological functions are controlled by 'modules', consisting of many types of molecules [12].

The notion of each module is that it is responsible for a separate, discrete function. The functions that these modules undertake cannot be easily predicted by studying the properties of the individual molecules that they are comprised of, thereby supporting the 'overall' view, as opposed to the individual molecular-basis one. However, identification of these modules is not always straight-forward, it is not always apparent what molecules they are made up of.

The idea that biological functions are carried out by various molecules is one of the reasons that there has been a shift away from reductionism towards systems biology. Due to the isolated approach of reductionism, dynamic interactions between parts are disregarded. Instead the human body is depicted as a collection of static biological components, with the emphasis placed on static stability, as oppose to dynamic stability. If we consider that biological functions are carried out by a number of molecules by modules, then it is clear that taking a systems approach over a reductionist approach is beneficial. The table below highlights the differences between the two approaches [2].

TABLE 2.1 DIFFERENCES BETWEEN REDUCTIONIST AND SYSTEMS APPROACH

CHARACTERISTIC	REDUCTIONIST APPROACH	SYSTEMS APPROACH
PRINCIPLE	BEHAVIOUR OF A BIOLOGICAL SYSTEM CAN BE EXPLAINED BY THE PROPERTIES OF ITS CONSTITUENT PARTS	BIOLOGICAL SYSTEMS POSSESS EMERGENT PROPERTIES THAT ARE ONLY POSSESSED BY THE SYSTEM AS A WHOLE, AND NOT BY ANY

		ISOLATED PART OF THE SYSTEM
METAPHOR	MACHINE, MAGIC BULLET	NETWORK
APPROACH	ONE FACTOR IS SINGLED OUT FOR ATTENTION AND IS GIVEN EXPLANATORY WEIGHT ON ITS OWN	MANY FACTORS ARE SIMULTANEOUSLY EVALUATED TO ASSESS THE DYNAMICS OF THE SYSTEM
CRITICAL FACTORS	PREDICTORS/ASSOCIATED FACTORS	TIME, SPACE, CONTEXT
MODEL CHARACTERISTICS	LINEAR, PREDICTABLE, FREQUENTLY DETERMINISTIC	NON-LINEAR, SENSITIVE TO INITIAL CONDITIONS, PROBABILISTIC
MEDICAL CONCEPTS	HEALTH IS NORMALITY, RISK REDUCTION AND HOMEOSTASIS	HEALTH IS ROBUSTNESS, ADAPTION AND HOMEODYNAMICS

One of the most revolutionary developments in recent biology has been the Human Genome Project. A fundamental idea emerged from this, the view of biology as an informational science [13]. Other sciences such as chemistry, physics, and geology, are measured through observations based on analogue results; at the heart of biology lies the genome, a digital code. The nature of digital codes, being discrete rather than continuous data, means that biology lends itself particularly well to disciplines such as computer science and engineering.

2.2 Biological Networks

High-throughput data techniques have been developed, such as the use of yeast two-hybrid screens and protein chips, that allow for integration of the components of the cell, and provide information about the status of molecular interactions within the cell [14]. One of the key concepts in systems biology related to this is the use of biological networks [15]. Biological networks are abstract representations of biological systems that capture a number of the essential characteristics that make up that system; the nodes represent molecules, and links represent interactions between the molecules. These networks provide a first inkling of the overall structure of molecular interaction networks of biological systems. Whilst the focus of this thesis concerns GRNs, it is worth introducing protein-protein interaction (PPI) and metabolic networks at this point.

PPI networks model protein binding interactions in an organism, and in humans provide a valuable tool to better understand the functional organisation of the proteome [16]. The first PPIs were generated using two-hybrid studies in the yeast organism Saccharomyces Cerevisiae [17]. Other organisms, such as humans [18] and Drosophila [19], have been the subject of protein interaction studies using large-scale two-hybrid studies. More recently, Xu and Li [20] used topological properties to discover hereditary disease genes in a human PPI network, highlighting the application of graph theory metrics, introduced in a later section, to PPI networks.

Inside an organism, chains of reactions, known as metabolic pathways [21], combine to perform particular functions, such as the process of glycolysis, extracting energy from food. Nodes represent compounds, and edges represent reactions in a metabolic pathway. The system of connected metabolic pathways is the metabolic network of the organism [22]. There are a number of repositories that detail metabolic pathway information for an organism, such as KEGG [23], facilitating the construction of metabolic networks for organisms. As with PPI networks and GRNs, metabolic networks can be analysed using graph theory concepts. These graph theory concepts are introduced in a later section of the thesis.

GRNs are the main biological network used in this thesis, and are used to model the interactions between genes in an organism. In a GRN, RNA transcripts of genes are represented as nodes, and relationships between these are represented as links [24]. These edges represent presumptive relationships between the RNAs; the amount of one RNA affects the amount of the other RNA. These relationships can be simple, where RNA C encodes a transcription factor that promotes the transcription of RNA D, or complex where multiple metabolites or protein signalling is involved. Whilst these metabolites and proteins are not explicitly shown in the GRN, they may be involved in the relationship shown in the link between two genes.

In this thesis, GRNs are constructed from microarray data from cancer cell lines; GRNs can also be constructed from other types of data, such as RNAseq experiments. To date, a number of GRNs have been constructed for various types of cancer; these include prostate cancer [6], breast cancer [7], and brain cancer [7]. Microarrays, and GRN inference methods, will be explained and discussed in detail in the next chapter.

Despite the huge amount of data available, very few biological networks are complete in their structure [25]. This incomplete structure has given rise to the prevalence of mathematical models that represent the different biological networks. Historically, the process of evolution has been viewed by some, such as François Jacob [26], as being somewhat haphazard, with parts being added or taken away until a working solution is found. However, recent advances in the understanding of biological networks, using engineering concepts, have found that the solutions created by evolution share a number of features with good engineering design [27], a perhaps somewhat surprising result. Three main engineering principles have been identified in biological networks; modularity, robustness to component tolerances, and the use of recurring circuit elements.

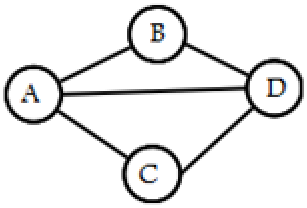
The idea of modularity has already been noted in the previous section, that biological functions are controlled by ‘modules’ of molecules that work together. This is further illustrated by the example of proteins that function in co-regulated overlapping groups such as pathways [28]. Modules are apparent in engineered systems; such as subroutines in software [29]. Robustness to component tolerances is a well-known and observed feature of engineered systems; the system should function under all conceivable conditions that may arise due to the components and the surrounding environment. This has also been observed in biological network, recent studies have shown how particular gene circuits are robust in bacterial chemotaxis [30] and fruit-fly development [31].

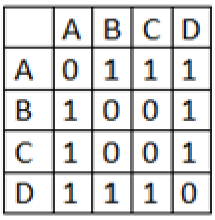
The third observed principle is the use of recurring circuit elements. Electronic devices, such as computers, include thousands of circuit elements such as memory registers. This same principle is apparent in biological design; key wiring patterns are repeated throughout a network. One such example of this is the presence of recurring ‘network motifs’ circuit elements observed in the transcriptional network of E.Coli [32]. In order to analyse biological networks, and more specifically GRNs, tools and concepts from the field of graph theory can be used. Graphs can be used as representations of real world systems. These tools can identify important genes in the GRN based on topological properties. Relevant graph theory tools and concepts are introduced in the following section.

2.3 Undirected Unweighted Graph

In the field of mathematics, the graph, G, consists of vertices, V, that represent points, and edges, E, that represent links between the points and can be defined as (V,E) [33]. In essence, a graph is a series of points connected by a series of links. A connection between nodes i and j is defined as $E = \{(i, j)|\ i, j\ \hat{I}\ V\}$. As there is an edge between nodes i and j, they are said to be neighbours. It is possible for more than one edge to exist between nodes. The simplest graph model is the undirected unweighted graph, no direction is assigned to the edges, and no weighting either, so all edges in this type of graph have the same strength. A graph can be represented mathematically using an adjacency matrix. The graph containing n vertices would be represented by an n *x* n matrix $A = (a_{i,j})$; with the entry $a_{i,j} = 1$ to indicate an edge connecting vertex i to vertex j, and $a_{i,j} = 0$ if there is no connection from vertex i to vertex j [34]. Note that for undirected graphs, the adjacency matrix is symmetrical, as no distinction is made between the origin and destination of the edge. An example of an undirected graph and corresponding adjacency matrix is shown in figure 2.1 below:

FIGURE 2.1 UNDIRECTED UNWEIGHTED GRAPH AND CORRESPONDING ADJACENCY MATRIX



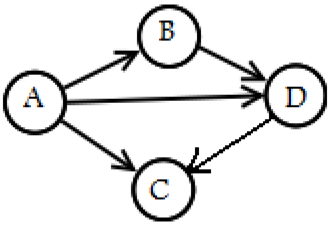


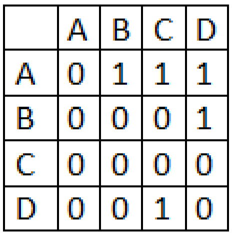
	A	B	C	D
A	0	1	1	1
B	1	0	0	1
C	1	0	0	1
D	1	1	1	0

2.4 Directed Unweighted Graph

In a directed graph, the edges have directionality. This builds upon the previous model, and directionality is assigned to the edges. This captures the directionality that exists in certain real world situations; such as communications networks, where there is a sender and a recipient of a message. In this case, an edge originating at i, representing the sender of the message, and finishing at j, the recipient of the message, would be represented as $E = (i, j)$, and not $E = (j, i)$ [33]. As with the undirected graph, an adjacency matrix can be used to represent it mathematically; with a 1 indicating an edge, and a 0 representing no edge. Note that this matrix is not symmetrical due to the directionality of the edges. This can be seen on the example of a directed graph and corresponding adjacency matrix is shown in figure 2.2 below, there is an edge from vertex A to vertex B, not from vertex B to vertex A.

FIGURE 2.2 DIRECTED UNWEIGHTED GRAPH AND CORRESPONDING ADJACENCY MATRIX





	A	B	C	D
A	0	1	1	1
B	0	0	0	1
C	0	0	0	0
D	0	0	1	0

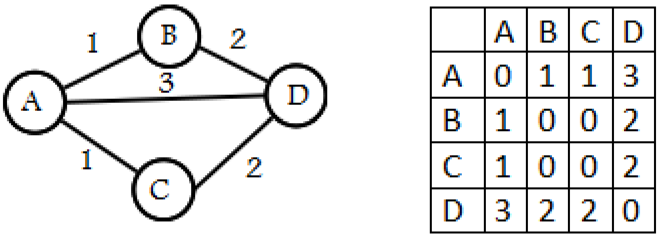
2.5 Undirected Weighted Graph

Another extension to the first model introduced is the assignment of weights to the links that connect the nodes. The weighted graph differs from the unweighted graph in that edges between nodes are not discrete in nature, either 1 or 0, but instead edges have respective weights that show the strength of the edge in relation to the other edges in the network. There are a number of scenarios where an edge weight is useful for providing additional information. One such example of this is a map represented as a graph; the nodes represent cities, the edges routes between cities, and the edge weight distances. Weighted networks can be represented mathematically by an adjacency matrix, as is the case with unweighted networks, but instead of having entries that are binary, 1 or 0, instead these entries are equal to the relative weights of the edges [35]. The adjacency matrix A therefore has elements:

Aij = (weight of connection from i to j)

This can be further seen in the example undirected weighted graph along with corresponding adjacency matrix in figure 2.3 below:

FIGURE 2.3 UNDIRECTED WEIGHTED GRAPH AND CORRESPONDING ADJACENCY MATRIX

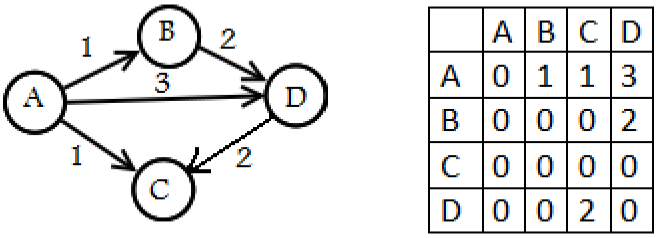


	A	B	C	D
A	0	1	1	3
B	1	0	0	2
C	1	0	0	2
D	3	2	2	0

2.6 Directed Weighted Graph

The fourth model adds directionality to the weighted graph, so that edges are both directed and weighted. As with the previous three types of graphs, it can be represented using an adjacency matrix. Figure 2.4 shows an example of a directed weighted graph with corresponding adjacency matrix:

FIGURE 2.4 DIRECTED WEIGHTED GRAPH AND CORRESPONDING ADJACENCY MATRIX



	A	B	C	D
A	0	1	1	3
B	0	0	0	2
C	0	0	0	0
D	0	0	2	0

2.7 Graph Models

The use of graphs as a mathematical modelling technique for real world systems is not a new phenomenon, and can be traced back to the 1730s. Leonard Euler, a Swiss mathematician, used it to solve the bridges of Konigsberg problem [36]; was it possible to walk across all seven bridges of the city, and never cross the same bridge twice? By modelling the bridges of the city as graph, Euler was able to prove that such a path did not exist due to four of the nodes have an odd number of edges. When a new bridge was built in 1875, such a walk finally became possible.

The number of edges that a node has can be thought of as perhaps the most basic quantitative property of a graph, and is termed the degree. The degree of a node is the number of edges that are connected to that node [37]. The distribution of the number of connections each edge has is therefore the degree distribution. In 1957, the first serious attempt to construct a model for large and apparently random networks, the random graph, was proposed by Rapoport et al [38], and rediscovered independently a few years later by Erdős and Rényi [39]. Extremely simple models of networks were proposed; take n number of vertices and connect each pair with probability p or 1-p, resulting in the model having a clear Poisson degree distribution. This model demonstrated the most important property of the random graph, the transition from low-density, low p-state to high density, high p-state.

However, these early models had two major shortcomings. Firstly, they did not capture the property of clustering. This property, also known as network transitivity, is the observation that two vertices that are both neighbours of a separate third vertex have an increased probability of also being neighbours themselves [40]. It is more likely that two of your friends will also themselves be friends with each other, than if they were not friends with you. This property can be quantified by the clustering coefficient:

[2.1]

On a fully connected graph, this number is 1, in real-world networks the value tends to be between 0.1 and 0.5 [40]. Due to the random and independent probability of two nodes being connected in a random graph, random graphs exhibit a low clustering coefficient.

Watts and Strogatz [41] proposed the small-world model as a graph generation model that generated graphs with a high clustering coefficient. By taking a standard ring lattice, containing n nodes each with k edges, and rewiring the edges with probability p, they observed how altering this value of p affected the clustering coefficient. This is the foundation of the small-world model, a network built on a low-dimensional regular lattice with edges added or moved to create low density shortcuts that join remote parts of the lattice to each other. As well as generating graphs with a high clustering coefficient, the graphs also have a small average path length, both small-world properties.

Arguably the most famous discovery in the study of networks, popularised by science books and social networking experiments, was the small-world effect; that most pairs of people are connected by at least one, and probably many, short chains of acquaintances. Whilst originally this applied to social networks, as highlighted by Milgram [42], it is not just confined to these, and seems to apply to many other networks. The connectivity of the Internet, gene networks, the power grid of the western United States, and the neural network of the worm *C.Elegans* all display small-world behaviour [41].

The average path length in the small-world model has received a lot of attention. No exact solution exists for this metric, denoted ℓ, however a number of partial exact results are known, as well as a number of approximate solutions for its behaviour. Small average path lengths have been observed in real-world scenarios; the average mean path length for academic co-authorship with Erdős is 4.65 [43]; whilst in Milgram's original study it was 4.4, 5.4, and 5.7 for the three groups of participants [42].

Despite generating graphs that encapsulate a high clustering coefficient and small average path length, the model proposed by Watts and Strogatz has one major flaw. The degree distribution generated by the model is not realistic; it does not match the degree distribution observed in real networks. This is also a shortcoming of the random graph model. Real-world networks do not have a Poisson degree distribution; Barabási and Albert observed that the world wide web was scale-free, that it followed a power-law degree distribution [44]. This power law states that the fraction of nodes in the network, *P(k)*, that have *k* connections to other nodes in the network is:

[2.2]

with *c* being a normalisation constant, and the parameter *y* typically having a value between 2 and 3 [45]. Scale-free networks are dominated by a small number of highly-connected nodes, referred to as hubs. A number of studies have noted that biological networks display this scale-free topology, such as those by Jeong et al [46] ,on the topology of metabolic networks, with authors of other studies proposing that the scale-free topology of biological networks is more conserved than content during evolution [47]. Zhang and Horvath [48] are even more explicit, stating that a number of biologists would be wary of gene correlation networks that did not display scale-free behaviour.

One of the properties that arises due to power-law degree distributions is the high resilience to uniform random removal of nodes; when nodes are removed from a network with a power-law degree distribution uniformly at random, the network typically remains connected and functional regardless of the number of nodes removed [49]. However, whilst scale-free networks are incredibly resilient against the random removal or failure of nodes, they are highly susceptible to targeted removal of the highest-degree nodes. It has been suggested that regardless of the power-law exponent, no more than 3 % of these nodes need to be removed before the entire network is disconnected [50], meaning the average probability of there being a path connecting any two nodes disappears.

2.8 Network Metrics

A number of metrics can be used to analyse the networks modelled by graphs. The most basic of these, the previously introduced degree, is simply the number of connections that a node has. In the case of weighted networks, where all the nodes in the network are connected to all the other nodes in the network, all nodes have the same degree. Therefore, the degree centrality property can be extended for weighted networks, by taking the edge weights into consideration. The sum of the edge weights of a node is of interest, and is referred to as the weighted degree centrality. The greater the strength of a connection, i.e. the weight of an edge between two nodes, the more informative this connection is likely to be. This is especially relevant in the case of GRNs; higher edge weights are indicative of a stronger regulatory relationship between two genes [51].

In addition to the number and strength of connections a node in a network has, it can be useful to monitor communications between all nodes in the network. More specifically, it is of interest to measure the effect that a particular node has on these communications. This can be measured using the betweenness centrality; it measures how many shortest paths between all the node pairs in a network pass through a specific node [52]. The shortest path in a network is defined as the path that requires the least amount of intermediary nodes between a pair of nodes, and can be defined formally as [53] :

[2.3]

h=intermediary nodes on path between nodes i and j

Each node that is part of the shortest path between a pair of nodes is able to control the information that flows between these nodes. As such, a node that is part of many shortest paths is important in the network as it controls the communications between many nodes in the network [54].The formal definition of betweenness centrality, as given by Freeman, is [55]:

—— [2.4]

=number of shortest paths between two nodes, and number of those paths that pass through node i

A node with a high betweenness centrality in a network is crucial to maintaining certain connections, and its removal can result in the network becoming disconnected; there no longer is a path in the network between certain nodes, resulting in the nodes no longer being able to communicate with each other. In a biological context, Scardoni states that the betweenness of a node in a protein signalling network gives an indication of the relevance of the protein as capable of holding together communicating protein [56]. A study by Potapov et al [57] on mammalian transcriptional networks identified the betweenness centrality as probably being the most biologically significant topological property, indicating that a gene with high betweenness is more likely to be involved in regulatory mechanisms.

However, in the case of a weighted network where, all nodes are connected to other nodes, the above implementation of betweenness centrality is not applicable. As such, a modified implementation taking into account the edge weights is required. There have been a number of different approaches to the area of identifying shortest paths in a weighted network. Dijkstra [58] suggested an approach based on the path of least resistance, i.e. prioritising paths that involved lower edge weights. This approach is suited to networks where the edge weight represents a cost, such as time or money; as would be the case for a network representing a city metro system, where edges weights are the costs of travelling between stations. However, in a number of networks, the edge weight represents a measure of optimality, and as such, paths with higher edge weights should be prioritised over paths with lower edge weights. Both Newman [59] and Brandes [60] inverted the edge weights in order to achieve this. The formal definition of their implementation of the shortest path algorithm is then:

$$— \qquad — \qquad [2.5]$$

Another well-studied network metric is that of closeness centrality. It is of interest to be able to measure how quickly a node can reach all the other nodes in the network. The closeness centrality is the inverse of the average length of the shortest paths to/from all the other vertices in the graph, and is defined as [55]:

$$C_c(_i) = \left[\sum_{j=}^{g} d\ _{i,j}) \right] \qquad [2.6]$$

A node with a high closeness centrality will have a small average distance to all other nodes in the network. Scardoni [56] states that in a protein-signalling network, a protein with high closeness will be central to the regulation of other proteins; but that some of the proteins will not influenced by its activity. It might therefore be expected a gene with high closeness in a GRN is likely to be involved in some regulatory mechanisms, but will not have any influence on others; although this has not explicitly shown.

As was the case with betweenness centrality, we are interested in extending this metric to weighted networks. Newman extended closeness centrality to weighted networks using a similar approach adopted for extending betweenness to weighted networks detailed previously

[59]. The edge weights in the network were inverted, and the least costly paths for all pairs of nodes were found. This cost of the path indicated how far a node was from another, and as such, the higher the cost, the further away a node was. In order to then create the closeness measure, this cost was inverted, so that high costs were transformed into a low closeness centrality, and low costs were transformed into high closeness centrality [53].

Another network metric of interest is the eigenvector centrality. This metric, initially proposed by Bonacich [61], can be thought of as an extended version of degree centrality. Scores are assigned to nodes based on the sum of the degree of the nodes they are connected to; therefore a node can have a high eigenvector centrality either by being connected to lots of others nodes, or being connected to nodes that themselves have a high degree. The famous PageRank algorithm used by the Google search engine is a variation of the eigenvector centrality metric. The eigenvector centrality can be formally defined as [62]:

$$x_v = \frac{1}{\lambda}\sum_{t\in M(v)} x_t = \frac{1}{\lambda}\sum_{t\in G} a_{v,t}x_t \qquad [2.7]$$

= score of the vth node,

M(v)= set of nodes v connected to,

= adjacency matrix of the network i.e. = 1 if v is connected to t,

λ = constant

Re-writing this in matrix notation, we get:

$$\lambda x = Ax \qquad [2.8]$$

as a result we can see that x is an eigenvector of the adjacency matrix. Extending the eigenvector centrality to a weighted network, we can take into account the edge weights. This means that if node B was connected to node A with double the edge weight of the edge from node A to node C, node B would contribute twice as much as node C to the eigenvector centrality of node A.

As noted earlier, typical network connections are far from random. One indicator of this is the correlation between degrees of different nodes; are highly connected nodes connected to each other? To formally measure this, the assortative mixing coefficient of the network can be calculated. This is the Pearson correlation coefficient of nodes that are connected. Newman [63] defines this for an undirected network as:

$$r = \frac{\sum_{jk} jk(e_{jk} - q_j q_k)}{\sigma_q^2} \qquad [2.9]$$

j,k=excess degree of the specific edge

= joint probability distribution of excess degrees of the nodes at either end of a randomly chosen edge,

= standard deviation of the excess degree distribution of the network q(k)

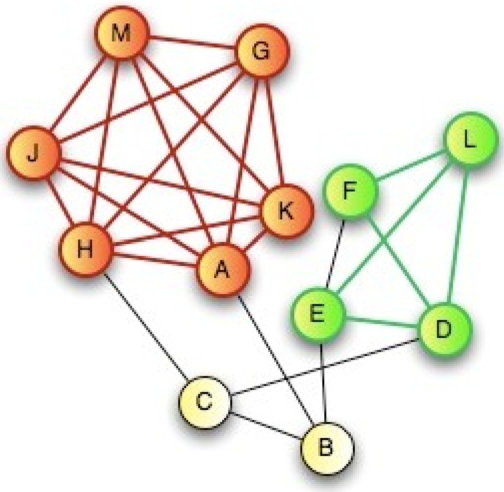
Where r = 1, there is perfect assortative mixing, where r = -1, there is perfect diassortative mixing. In social networks, there are positive correlations, r has been observed to be positive, meaning that the highly social party animals tend to know and socialise with each other. However, most non-social networks, including biological networks, have negative correlation. These degree correlations have a strong effect on the structure of networks. Networks with a positive correlation have a core-periphery structure; a highly interconnected core surrounded by periphery lower degree nodes, networks with a negative correlation have high-degree nodes scattered more broadly over the network. These structural differences also affect the way a network behaves, for example a disease can persist more easily in a network with positive correlation by circulating in the dense core where there are opportunities for it to spread, in a network with negative correlation it is harder for the disease to persist, but if it manages to do so, it will typically spread to the whole of the network.

The greatest distance between any pair of nodes in a graph is the diameter. It is relatively straightforward to calculate; calculate all the shortest paths in the graph as per the betweenness centrality, the largest length of these paths is the diameter. The diameter can give an initial insight into how compact a graph is; it is possible that two nodes are far away from each other in a graph giving a high diameter, yet the graph itself is quite compact. This should be taken into consideration before making conclusions about how compact a graph is based on diameter. Low diameter values have been observed generally in biological networks [64], a low diameter in a GRN can indicate that the genes are able to communicate with each other relatively easily.

Cliques are another area of interest. The clique, also called the complete graph, is a complete subset, *S*, of a graph in which each pair of nodes is connected. In particular, the maximum clique is of interest; this is the largest clique that exists in the graph. The maximum clique can be informative in a GRN as it highlights gene co-expression relationships, i.e. genes up-regulated or down-regulated together. It has been shown that genes that form part of this maximum clique are likely to be functionally related, and this can give an insight into cellular processes [65].

The maximum clique containing a specific set of nodes can also be informative. The maximum clique containing nodes *F, D,* and *L,* shown in green in figure 2.5, also includes node *E*. This indicates that node *E* is more likely to be part of cellular processes that nodes *F, D,* and *L,* are involved in [65]. The clique coloured red in figure 2.5 is the largest clique in the graph.

FIGURE 2.5 CLIQUES IN A NETWORK



2.9 Summary

This chapter has introduced a number of theoretical concepts that are central to the thesis. An introductory overview of systems biology was presented, along with the notion of biological networks. GRNs in particular were detailed, with examples of their application to cancer cell line data sets. Relevant concepts of graph theory relating to the thesis were outlined; different types of graphs, mathematical models of graphs, and a number of properties of graphs. In particular, graph theory metrics were presented that can be used to analyse GRNs, and the context of their application to GRNs. In the next chapter, the specific data type that will be used for the thesis is briefly explained, along with a number of potential GRN network inference methods, and a proposed methodology for the thesis.

Chapter 3

Gene Regulatory Network Inference Methods and Proposed Methodology

In this chapter, a brief introduction to microarrays is presented, along with a description and discussion of various GRN inference methods. Finally, the proposed method for this thesis is presented.

3.1 Microarray Data

For the purposes of this project, GRNs will be inferred from microarray data. Before different GRN inference methods are detailed, microarray data will be introduced. In order to understand microarray data, the underlying biology has to be first understood. Almost all of the cells in the body contain a set of chromosomes and identical genes, with only a small portion of these genes turned on. It is this small fraction of genes that are expressed and that are responsible for the unique properties of each cell. The term gene expression refers to the transcription of information contained in the DNA into messenger RNA, mRNA, then translated into proteins responsible for the critical functions of most cells. By studying the type and amount of mRNA produced by a cell, we can get an insight into how the cell behaves. The process of gene expression is highly complicated, and allows a cell to respond to the environment and its own needs. As well as genes turning on and off, the level of expression can be increased or decreased. The correct expression of a number of genes is required for normal growth and development; changes to these expression levels are the cause of a great deal of diseases.

A microarray enables the gene expression levels of thousands of genes to be measured in one experiment. It works by exploiting the ability of a mRNA to bind to the DNA template it originated from; the mRNA is fluorescently labelled and the fluorescence is measured by a scanner generating the gene expression level [66]. It is these gene expression levels that then make up the array which is used for the inference of the GRNs.

It is necessary for a number of procedures to be carried out on an array before GRNs can be inferred from it. Essentially, three steps have to be carried out on a microarray before it can be used; background correction to remove background signal, normalisation to correct for systematic biases such as different dye absorption, and a summarisation of the values of all the probes representing one gene [67]. RMA, Robust Multichip Analysis, is the most common pre-processing method applied to microarray data, as it is performs all three steps [68]. Additionally, in studies such as those by Freudenberg [69] and Hill et al [70], RMA has been shown to outperform other pre-processing methods, including the Affymetrix MAS5 algorithm.

3.2 GRN Inference Methods

In order to infer a GRN from microarray data, a number of different approaches can be used. We will focus on four techniques that can be used; mutual information based methods, correlation based methods, Bayesian networks, and a discretised approach.

3.2.1 Mutual Information Networks

The first category of network inference model that will be introduced is the mutual information based approach. Mutual information is an indicator of being able to predict the value of one variable, knowing the value of the other variable, and is defined mathematically as [71]:

$$M(X,Y)=\sum_{y\in}\sum_{x\in}\left(x,y\right)\log\frac{p\left(x,y\right)}{p(x)p(y)}\qquad[3.1]$$

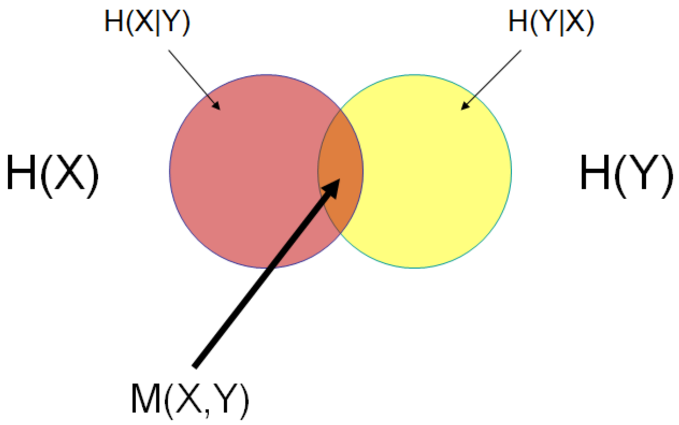
where p(x,y) is the joint probability of variables X and Y,

p(x) is marginal probability distribution function of X,

p(y) is marginal probability distribution function of Y

An illustrative way to display mutual information is on a Venn diagram, shown in figure 3.1:

FIGURE 3.1 VENN DIAGRAM DISPLAYING MUTUAL INFORMATION



$$M(X,Y)=H(X)-H(X\,|\,Y)$$

The mutual information network inference approach uses mutual information to determine if a link exists between two nodes in a network. There are two steps involved in the inference of a network using mutual information. The first step is the calculation of the mutual information matrix. This is a square matrix, where:

$$MIM_{ij} = I(X_i{:}X_j)\qquad[3.2]$$

represents the mutual information between X_i and X_j [72]. Having calculated the mutual information matrix, the next step is to then use an inference algorithm to calculate an edge score for each pair of nodes in the network, using the mutual information matrix. One of the advantages of mutual information methods should be highlighted here; they are able to deal with thousands of variables with only a limited number of samples, making them ideally suited to microarray data. It should be noted that as the mutual information is a symmetrical measure, edge directionality cannot be inferred using this method.

A number of inference algorithms for calculating the edge scores from the mutual information matrix have been developed. Amongst the most promising for use with microarray data to infer GRNs are ARACNE [73] and MRNET [71]. ARACNE in particular has received a lot of attention. It is an inference algorithm based on the Data Processing Inequality; if gene X_{A} interacts with gene X_{C} through gene X_{B}, then [74]:

$$I(X_{\mathrm{A}}; X_{\mathrm{C}}) \leq \min\left(I(X_{\mathrm{A}}; X_{\mathrm{B}}), I(X_{\mathrm{B}}; X_{\mathrm{C}})\right) \qquad [3.3]$$

Each pair of nodes is initially assigned a weight equal to the mutual information matrix by ARACNE. A threshold I_0 is set, and all edges which are below that threshold, i.e. $I(Xi; Xj) < I_{0,}$ are removed. ARACNE interprets the weakest edge of each triplet as an indirect interaction, and this is removed if the different between the two lowest weights is greater than another threshold, W_0. In the study outlining its use, ARACNE was shown to perform better than relevance networks and Bayesian networks on a number of network inference tasks [73].

MRNET is a network inference algorithm that uses the maximum relevance/minimum redundancy (MRMR) feature selection method [75]. This approach scores the set of inputs, V, based on the difference between the mutual information and the output, Y, the maximum relevance, and the minimum redundancy, which is the average mutual information of the previously ranked variables. The idea is that this approach will highly rank direct interactions, whereas indirect interactions will be given a low ranking. MRNET performs similarly to ARACNE, and outperforms it for precision recall from simulated datasets [76].

3.2.2 Correlation Networks

Another network inference approach is to use correlation networks. Correlation networks are used in a wide variety of applications such as finance [77] , ecology [78], gene expression analysis [79], and metabolomics [80]. A correlation network displays the correlation that exists between two elements, with 1 representing a perfect positive correlation, and -1 a perfect negative correlation. The creation of a correlation network is in itself a relatively easy process consisting of two steps; the calculation of all pairwise correlations, and the application of a threshold to identify significant correlations that hence form edges in a network [81].

Gene correlation networks are a tool used to explore the role of genes in a system context. The concept of gene correlation networks is fundamentally simple; nodes represent genes, and edges represent interactions between these genes. A gene co-expression matrix can be defined:

$$S = [ij] \qquad [3.4]$$

where ij is the correlation between the genes i and j,

S is the matrix of correlations.

A common approach is to use the Pearson Correlation coefficient between pairs of genes; and to then assign either an edge being present, or absent, based on a threshold value, resulting in an unweighted adjacency matrix that can be used to construct a network [82]. A variation on this is to represent all the correlations between pairs of genes using a weighted adjacency matrix, although this results in a much more computationally demanding matrix.

A number of previous studies have used correlation based methods successfully for dealing with biological data, such as those by Stuart et al [83] , Li et al [84], and Horvath et al [7]. These results from these previous studies show the practical usefulness of the correlation based approach for dealing with microarray data. In particular, the weighted correlation network analysis (WGCNA) approach proposed by Horvath et al [85] has provided a number of interesting results, in particular when the method was applied to a glioblastoma dataset.

3.2.3 Bayesian Networks

Bayesian networks are another type of network inference method. Think of a situation where two events have an effect on a third event; for our example, a car is clean depends on whether it has been raining or whether someone has washed it. Whether the car is washed also depends on the rain; the car is not washed when it is raining. In such a scenario, we can use a Bayesian network to model this. Bayesian networks are probabilistic graphical models that represent probabilistic dependencies between a set of entities, and in doing so combine two areas of mathematics; graph theory and probability. The Bayesian network is a directed acyclic graph, $G(X, E)$, where gene expression levels are represented by random variables that are the nodes, $x_i \in X$, and the dependencies between these nodes are represented as edges [86]. The variables representing the nodes are derived from conditional probability distributions, $P(x_i|Pa(x_i))$, where $Pa(x_i)$ is each node's set of parents. The Markov Assumption, that each variable is independent of its non-descendants given its parents, is implicitly encoded by the Bayesian network. Building on this assumption, the joint probability distribution over all the variables down to the conditional distribution of the nodes can be defined as [24]:

$$P(x_1, x_2 ... x_n) = \quad (x_i|Pa(x_i)) \qquad [3.5]$$

As well as the set of dependencies implied, that children nodes are dependent on their parent nodes, a set of independencies are also implied.

The Bayesian network is appealing for modelling deterministic relationships, such as GRNs, as the joint probability distribution for the probabilities of events, represented by the nodes, given other events can be queried. Inferences can be made from the joint distribution and likely causal relationships can be made. The probability based nature of Bayesian networks also makes them suited to handling noise that is inherent in microarray data.

In order to use a probabilistic framework for a GRN, the aim is to learn a Bayesian network that is a good fit for the microarray data. There are two steps to learning a Bayesian network from the data; the first is the selection of the model, and the second is the parameter fitting [87]. The model selection consists of finding the best graph, *G,* to represent the relationship between the variables of the data. The second step then consists of finding the best conditional probabilities for each node of this model. A number of methods have been proposed for learning Bayesian networks from genomic data, such as those proposed by Werhli et al [88], and Chen et al [89], have proved promising for inferring GRNs from microarray data using a Bayesian approach.

3.2.4 Discretised Networks

A fourth conceptually much simpler network inference method is the discretised approach. Vass et al [90] proposed such an approach for inferring a GRN from microarray data. In their work, firstly the gene expression levels for all the genes are converted to z scores; and are then dicretised into 3 values based on a threshold value. The 3 discretised values are 1, which represents an up regulated gene, -1 which represents a down regulated gene, and 0 which represents neither, and are stored in a matrix. This matrix of discretised values is used to derive two matrices, P and M, which hold the 1 and -1 values in positive form. These matrices are then used to calculate PP, the positive – positive network, i.e. where up regulation in one gene occurs with up regulation with another gene, negative – negative, where down regulation in one gene occurs with down regulation of another gene, and positive – negative, where up regulation of one gene occurs with down regulation of another gene.

In order to filter out relationships that may be as a result of chance, all the scores for these matrices are evaluated against an expectation, P = 0.005, using Monte Carlo sampling. Monte Carlo methods can be used to solve various types of computational problems through the use of random numbers [91]; Vass et al estimated the distribution of scores for randomised vectors of all possible densities, and repeated this test 1000 times. Values which exceeded 99.5% of the random scores were accepted, resulting in 5% of original edges remaining.

This simple discretised approach captures whether genes are up and down regulated for certain samples compared to others, which is often what biologists are interested in. It was shown that this method identified most of the defined relationships in the synthetic microarray dataset created using SynTReN [92], as well as identifying a great number of gene-gene relationships in observational microarrays across different mRNA platforms.

3.3 Comparison of GRN Inference Methods

In the previous section, a number of inference methods that can be used to infer GRNs from microarray data were introduced. However, a great many more network inference methods exist, and it would be impossible to introduce them all. Accurately reconstructing GRNs from microarray data is still a big challenge, despite the huge range of different network inference methods available.

Such is the challenge, that a project specifically dedicated to this problem exists; the dialogue on reverse engineering Assessment and methods (dreAm) project. This consortium meets annually to assess different network inference methods that have been proposed. At the 2010 conference [93], over 30 network inference methods were assessed, without one inference method identified as performing optimally across all of the datasets. From experience with the network inference methods used so far, all perform significantly better on their training set data than they do on other datasets, an observation that is in concordance with those from the 2010 dreAm conference.

Due to this large number of different inference methods, it is both impossible to use all of them, and also, to form a conclusive opinion as to which is the best-suited for the project. With this in mind, a number of network inference method review papers that have compared the respective performance of a number of different network inference methods have been referred to for initial guidance. These papers compared the performance of mutual information, correlation-based, and Bayesian network approaches. The discretisation approach proposed by Vass et al did not feature in any of the network inference review papers consulted.

The Bayesian network and discretisation approach potentially offer advantages over the two other approaches in that they have the potential to assign directionality to the interactions. However, in a number of studies, including that by Hurley et al [9], it was reported that the inferred directionality from a number of network inference methods, including Bayesian networks, was incorrect. Reasons for this could be due to the complexity of gene regulation;

gene A may regulate gene B, but both genes A and B are regulated by gene C and so on, and also there is limited directional information available in the microarray dataset for the method to infer the correct directionality. Whilst directionality offers a number of benefits to identifying genes of interest, if the inferred directionality is not correct, as the studies suggest, then this potential advantage of directionality is negated. As such, the networks that will be inferred will not be directed.

Another GRN inference methods study, by Allen et al [94], suggested that Bayesian networks are not suited to large scale microarray data due to computational and memory requirements. Learning the structure of a Bayesian network is NP-hard [95]. Considering that the datasets that will be used for this study consist of several thousand genes, Bayesian networks are not a feasible choice of network inference method for large-scale gene regulatory networks [96]. There is also high variability in the learning of the Bayesian networks; a slight difference in the parameter fitting and choice of model results in the inference of a completely different GRN. Taking these two negative aspects into account, Bayesian networks will not be used in this project, despite the potential advantages that they offer over other network inference methods.

It is worth noting that in the study by Allen et al, both WGCNA and ARACNE were identified as being a good choice of inference model for constructing the global network. These two methods also out-performed the other network inference methods when using the E.coli dataset, suggesting that they are also more robust than the other methods and also perform better with real datasets. The importance of the number of samples should be mentioned here. As with most other cases, increasing the number of samples will result in more meaningful results. This is especially pertinent in the inference of the GRNs from the data; a minimum number of samples are required for certain network inference methods. In the Allen el al paper, it was observed that as the number of samples increased, the performance of the network inference methods improved. Of particular interest was the performance of the network inference methods for the 20-50 samples category, in which WGCNA out-performed the others. This is of interest as the typical datasets that will be used contain this number of samples.

Finally, when considering the area of inferring a model that approximates a GRN from a microarray dataset, the words of David Edwards should be heeded: ‘Any method (or statistician) that takes a complex multivariate dataset and, from it, claims to identify one true model, is both naive and misleading’ [97].

3.4 Software Tools

A number of software tools will be used in this work; the main tool that will be used is the R programming language [98]. R is a language specifically developed for statistical computing and graphics, and builds on the S language developed at Bell laboratories. It has been widely adopted by bioinformaticians in recent years due to the wide range of bioinformatics libraries available; in particular the Bioconductor project [99]. Bioconductor project is an open source collaborative development project for the creation of software libraries in R for computational biology, and contains a wide range of software libraries with useful functions for applying to microarray data.

R will be heavily used in this work; the network inference methods to infer the GRNs from the samples will be implemented using R, as will the calculation of a number of different metrics for the GRNs and the identification of the largest unique cliques in the networks using the igraph [100] package in R. Igraph is package of functions specifically designed for complex network analysis. Filtering of the results will also be carried out in R, so that lists of highly ranked genes based on the metric scores, and the genes that comprise the largest unique cliques, are output for further analysis.

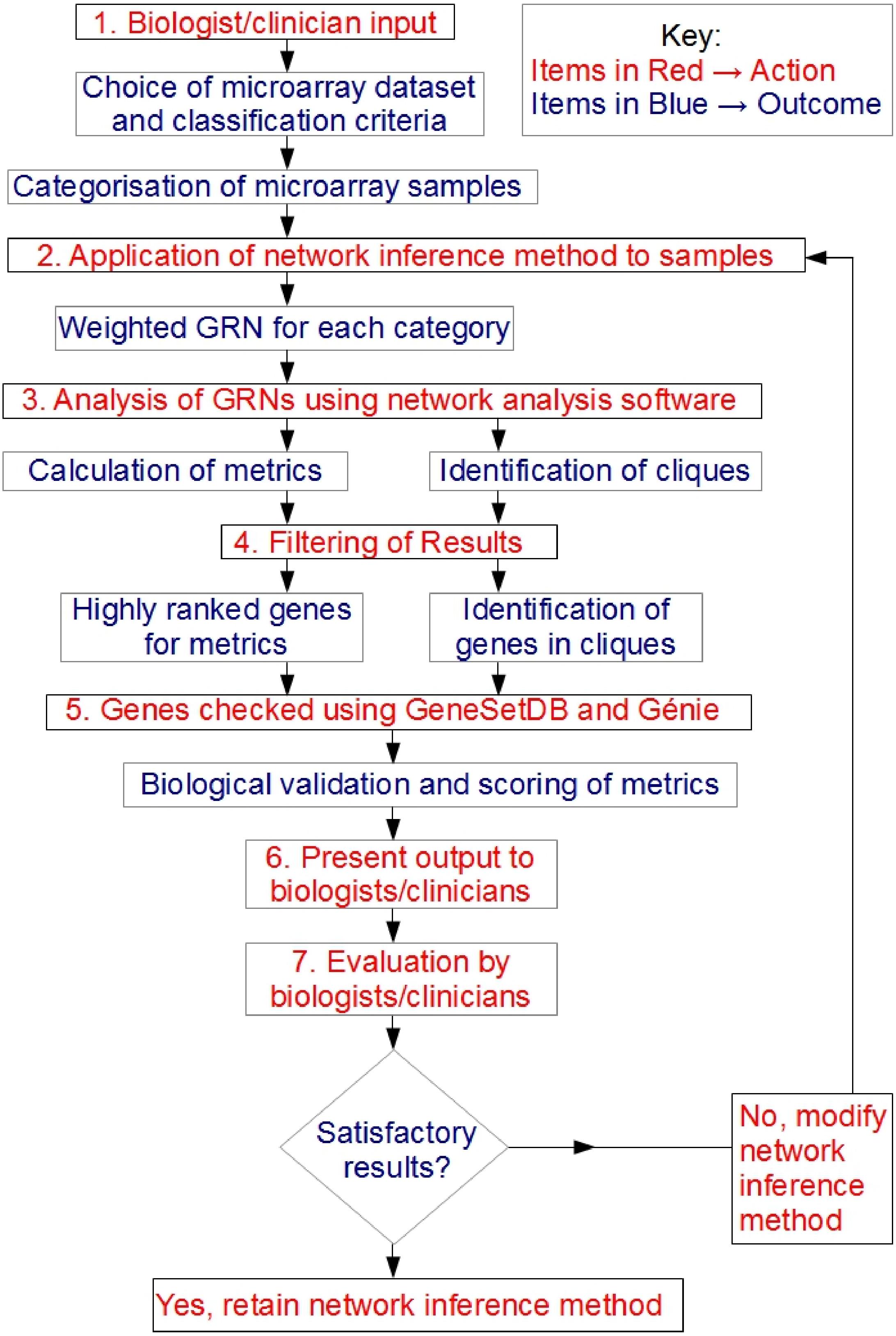
In order to counter the problem of assigning biological meaning to the lists of genes produced, there are a number of tools that are able to ascertain whether these lists have biological meaning. These tools highlight whether the genes identified are enriched for certain biological significance. Examples of widely used tools for this purpose include DAVID; Database for Annotation, Visualization and Integrated Discovery [101], a website with a number of tools for interrogation of the results, GATHER; *Gene Annotation Tool to Help Explain Relationships* [102], which is able to display annotations that distinguish the input list of genes from other genes in the genone, and GeneSetDB [103]; which can be thought of as an extended version of GATHER by using a greater number of databases, thereby ensuring greater coverage of the results.

The web-based tool Génie [11] will be used to score the genes identified. This tool scores genes for a selected biomedical topic based on a text-mining search of scientific abstracts. This thereby provides an objective means by which to scores genes for particular biological topics, rather than the subjective interpretation of GenesetDB that requires expert knowledge. For the purposes of this work, it provides a means to rank the metrics based on the scores of the genes they identify. For each microarray dataset, Génie will be used to generate a list of scored genes for the particular microarray disease.

3.5 Proposed Approach

Having reviewed the areas of interest that are applicable to this work, a proposed approach can now be presented to specifically address the aim and objectives outlined in chapter 1. The flowchart in figure 3.2 below shows the proposed approach that will be adopted.

FIGURE 3.2 OVERVIEW OF PROPOSED APPROACH


Key:
Items in Red → Action
Items in Blue → Outcome
1. Biologist/clinician input
Choice of microarray dataset and classification criteria
Categorisation of microarray samples
2. Application of network inference method to samples
Weighted GRN for each category
3. Analysis of GRNs using network analysis software
Calculation of metrics
Identification of cliques
4. Filtering of Results
Highly ranked genes for metrics
Identification of genes in cliques
5. Genes checked using GeneSetDB and Génie
Biological validation and scoring of metrics
6. Present output to biologists/clinicians
7. Evaluation by biologists/clinicians
Satisfactory results?
No, modify network inference method
Yes, retain network inference method

As can be seen from the flowchart above, there are seven steps to the proposed approach. The first thing to note is the role of the biologists and clinicians in the proposed approach in figure 3.2; initial input from these specialists will guide the selection and categorisation of microarray datasets that will be analysed; and once the network inference and analysis is complete they will be presented with the output lists of the genes and the enrichment results that will allow the results to be evaluated.

The first step of the proposed approach, the initial input from biologists and clinicians, will guide in ensuring that both the microarray datasets and classification criteria used are both relevant and of use to them, particularly as many biologists and clinicians are interested in certain diseases and subtypes of these diseases.

The second step is the application of a network inference method to infer the GRNs for the categorised samples. It is the intention of this work to explore the use of a number of GRN inference methods that can be used on microarray data; particularly to compare the use of an existing network inference method to infer a number of GRNs from a single microarray dataset, to the use of a novel network inference method to infer a number of GRNs from a single microarray dataset. Specifically, the performance of the WGCNA method, described in greater detail in the next chapter, will be compared to a novel network inference method, described in chapter 5, using a glioblastoma microarray dataset. This inference of GRNs from the different categories of samples directly addresses the first objective set out in chapter 1.

The third step is to analyse the GRNs inferred for the different categories. This directly relates to the second and third objectives set out in the first chapter. Metrics will be calculated for the networks using the igraph package in R, enabling genes to be identified for the networks based on these metric scores. The fourth step is to filter the results using R so that these genes are output for further analysis.

The fifth step is to interpret this list of genes using enrichment and gene scoring tools. For this work, the GenesetDB [103] tool will be used for enrichment. By using this tool, it is hoped that some biological insight into the results can be achieved. For the purposes of this work, the intention is to only use the top ten enrichment results returned by GenesetDB for each gene set. The analysis of the results is presented partly in this work to provide a demonstration of the ability of the network inference methods to derive results that are biologically meaningful for the microarray datasets they have been used on.

It should be noted that the output results of GenesetDB and the other enrichment tools detail specific biological processes that are hard for the casual user to interpret, thereby requiring specialists with expert knowledge to fully make sense of these. By presenting only the top ten enrichment results, this subjective need for expert biological knowledge is removed. In exceptional cases, other results outside of the top ten will be presented, but only when it is clear that they are related to the biological area of interest. Interpretation of the top ten enrichment results returned that will be presented in this work will be carried out using literature searches of the topics detailed. It should also be noted here that whilst GenesetDB has been chosen as the enrichment tool for this work, the biologists and clinicians may have their own enrichment tools that they feel are better suited to the purpose of interpreting the results. Again, this highlights the necessity for their expert interpretation and analysis of the results. Génie will be used to score the genes identified, providing an objective means by which to scores genes for particular biological topics, rather than the subjective interpretation of GenesetDB that requires expert knowledge.

The penultimate step is to present the results to the biologists and clinicians. A number of the tools that will be used to interpret the lists of genes make reference to the need for manual validation of the results, and it can be seen why this is the case. It should be remembered that the overall aim of this work is to aid biologists and clinicians in the interrogation of microarray data; whilst some analysis and interpretation is presented here, the full expert interpretation of these results is best suited to biologists and clinicians.

Related to this is the final step of the evaluation of the results that are obtained. It is important that the biologists and clinicians have confidence in the results, and also that the results make sense from a biological perspective. If the results do not follow basic biological principles, then in essence, they are not of use. By involving experts in the final evaluation, it is more likely that results will be produced that are meaningful.

3.6 Summary

This chapter has introduced a further number of concepts central to the thesis. A brief introduction to microarray datasets has been presented, from which the GRNs are inferred. A number of network inference methods for inferring GRNs have been outlined, along with a comparison of these different inference methods. Finally, the proposed approach that will be used in this work to address the objectives outlined in the first chapter has been outlined; including the introduction of the tool Génie for scoring lists of genes, and the tool GeneSetDB for biological interpretation of lists of genes. In the next chapter, an existing network inference method will be applied to a glioblastoma microarray dataset to infer a number of GRNs.

Chapter 4

Inference of Glioblastoma Survival Category GRNs using WGCNA

4.1 Introduction

In this chapter, an existing GRN network inference method, WGCNA, will be used to infer five different survival categories, based on survival time data, from a glioblastoma microarray dataset. These five GRNs will be analysed using the different graph theory metrics introduced in chapter 2. The text-mining tool Génie will be used to score the genes identified by the various metrics, and GenesetDB will be used to check for biological enrichment for the genes identified by each metric. This chapter will refer to a number of previous results from a study by Upton and Arvanitis [104].

The work in this chapter addresses the first two objectives outlined in the introductory chapter. The first of these concerns the implementation of a transferable GRN inference method to infer a number of GRNs from a single microarray dataset. The existing GRN inference method WGCNA is used here to investigate the feasibility of implementing an existing GRN inference method for this purpose. Some of the issues that arise from the application of an existing GRN inference method to infer multiple GRNs from a single microarray dataset are highlighted here. In the previous chapter, it was noted that in the study by Allen et al [94], WGCNA was the best performing GRN inference method when 20-50 samples were available. The number of samples available for GRN inference for each category of network for the dataset that will be used in this chapter roughly falls into this.

The second objective concerns the investigation of appropriate metrics to quantitatively analyse and compare GRNs constructed for different stages of disease, namely glioblastoma, in this chapter. The various graph theory metrics will identify genes of interest in each of the GRNs inferred, and these genes will be scored for their relevance to glioblastoma using the text-mining tool Génie. This allows an objective scoring of the genes that each of the metrics identifies in each category of glioblastoma GRN, and also allows the different categories of GRN to be compared. Génie does not rely on expert knowledge to interpret the results, something that is required when using the gene set enrichment tool GeneSetDB. Another possible approach to scoring the genes is based on their biological enrichment using GeneSetDB; however this requires expert biological knowledge to interpret the results. As such, the GeneSetDB tool will be used to check for biological enrichment, but will not be used in the process of scoring the genes. Furthermore, whilst references to the biological enrichment for the gene identified by the metrics will be made in this chapter and throughout the thesis, all GeneSetDB biological enrichment results will be presented in the appendix and not in main body so as not to overwhelm the thesis with tables, and the reader with specialist biological data that may not be relevant.

4.2 WGCNA GRN Inference Method

In this chapter, GRNs are inferred from a glioblastoma dataset of 120 samples from a previous study by Horvath et al [8]. The approach used by Horvath et al was to infer a single GRN from this dataset, using the WGCNA correlation-based network inference method. The main advantage of the WGCNA network inference method is that it is able to approximate the scale-free topology that is exhibited by biological networks [48]. Zhang and Horvath go further, and state that a number of biologists would be wary of gene correlation networks that did not display this scale-free behaviour [48]. Therefore, gene regulation networks inferred from microarray data should also observe this scale-free phenomenon.

Scale-free networks are approximated in WGCNA as a result of the method used to construct the weighted adjacency matrix. The weighted adjacency matrix used for WGCNA is constructed using two principal steps [85]. The first of these steps is to take the absolute value of the correlation between the nodes in question, in this case i and j, and define this quantity as the co-expression similarity as such:

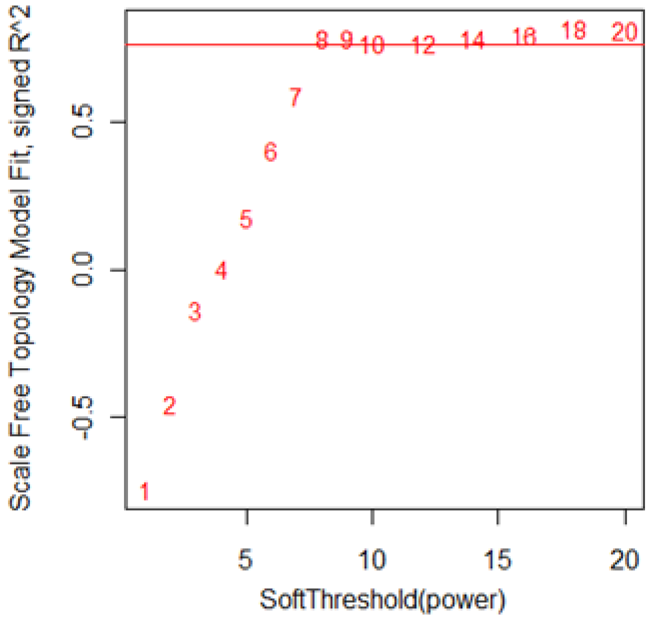
$$s_{ij} = |cor(x_i, x_j)| \qquad [4.1]$$

The weighted adjacency is subsequently created by raising the co-expression similarity by a soft threshold power, β, which is greater than or equal to 1, as such:

$$a_{ij} = s_{ij}^{\beta} \qquad [4.2]$$

This then allows the relevant power of *β* to be selected that allows scale-free topology to be exhibited by the network. As previously mentioned, the frequency distribution in a scale-free network is of the form . Inspection of whether a network fits the scale-free topology is achieved by plotting *$log_{10}(p(k))$* versus *$log_{10}(k)$*. If this plot is a straight line, it indicates that the network has a scale-free topology. Following on from this, the square of the correlation, , between *$log_{10}(k)$* and *$log_{10}(p(k))$*, can be used to indicate the relationship; if this value approaches 1, then there is a straight line relationship between *$log_{10}(k)$* and *$log_{10}(p(k))$*. A range of candidate values for the soft threshold power *β* can then be plotted, with the first value greater than 0.8 past the saturation point of the curve selected as the value of *β*. An example of this is shown in figure 4.1 below:

FIGURE 4.1 EXAMPLE PLOT OF SOFT THRESHOLD POWER β



From the plot above, it can be seen that a value of β= 8 satisfies the conditions to approximate scale-free topology, i.e. it is the first value greater than 0.8 past the saturation point of the curve.

4.3 Glioblastoma Microarray Dataset Categories

Glioblastoma is the most common primary malignant brain tumour in adults and also one of the most lethal forms of cancer, with a median survival time of only 14 months from the time of first diagnosis [105]. There are a number of previous glioblastoma studies, such as that by Verhaak et al [106], which have identified four glioblastoma subtypes, and a number of signature genes that correspond to each of these subtypes. This particular study found that the four subtypes had a narrow median survival range, from 11.3 months for the most lethal subtype, to 13.1 months for the least lethal subtype. Whilst identifying genes that corresponded to each subtype, this study did not identify genes that are specifically associated with glioblastoma prognosis, i.e. survival time. Therefore, in this chapter, the focus is on identifying genes that are associated with different survival times in glioblastoma cell lines, and can therefore be used as potential prognostic biomarkers of the disease.

The glioblastoma dataset consists of two independent sets of clinical tumour samples, 55 and 65 samples, respectively, obtained at the time of surgery at UCLA. Gene expression profiling of these samples was carried out using Affymetrix high-density oligonucleotide microarrays [107]. This dataset contains survival information, allowing the samples to be categorized based on survival time. A number of statistics based on this information can be calculated. The dataset as a whole has a mean survival time of 447.29 days, which correlates closely to the previously mentioned 14 month median survival time of glioblastoma. Median survival time of 336 days for this dataset is approximately 3 months less however. The standard deviation is 426.15 days.

These statistics highlight two observations; that the close correlation of mean and median survival time of the dataset with the clinical median survival time of 14 months indicates that this is a typically representative glioblastoma dataset, and also the high standard deviation value, which is very close to the mean, shows how survival time in the dataset greatly varies. In fact, the dataset has a range of 2800 days, with the lowest survival time of 7 days, and the highest of 2807. This gives an idea that potentially there are prognostic differences in the samples in the dataset, and that grouping all the data together into one single network to model the gene regulatory interactions in the dataset could lead to a great deal of information being lost about how the cancer evolves.

Five categories of samples were created from the glioblastoma dataset, using survival time as the class discriminator. Table 4.1 shows the five categories, along with the number of samples in each category and the mean, median and standard deviation.

TABLE 4.1 CATEGORIES, NUMBER OF SAMPLES AND STATISTICAL INFORMATION OF GLIOBLASTOMA MICROARRAY DATASET

Category	***Number of Samples***	***Mean (days)***	***Median (days)***	***Standard Deviation (days)***
200 days or fewer	37	112.11	112	54.37
201 to 400 days	35	298.66	302	59.25
401 to 600 days	18	500.22	500	72.05
601 to 800 days	15	677.98	667	53.57
800+ days	15	1326.73	1098	538.42

The first observation to be made is about how well spread out the samples are across the categories. It can be seen that the first two categories have a significantly higher number of samples than the other three. Initially, five categories were created with 24 samples each, although this approach was abandoned as the samples in each category were not so evenly spread out in terms of survival time. As can be seen, the 1st, 2nd and 3rd categories split well, as shown by mean and median being very close to the mid-point of each category. The next two categories are not so evenly spread out, and perhaps ideally could be further split; however from observations of the WGCNA algorithm, a minimum number of 15 samples are required for network construction. It is worth remembering the constraints that number of samples put on network inference, as previously mentioned, the greater the number of samples, the better the network inference method performs.

One analysis that highlights the value of the network inference method on the categories of data is to rank the mean expression for each gene in the five categories, and see how well these rankings correlate across the five categories. This gives an idea of how homogenous the data is across the five categories before the networks are constructed. Table 4.2 below shows the correlations.

TABLE 4.2 CORRELATION SCORES OF AVERAGE GENE EXPRESSION RANKINGS IN GLIOBLASTOMA CATEGORIES

Category	200 days or fewer	201 to 400 days	401 to 600 days	601 to 800 days	801 days and more
200 days or fewer	1	0.989	0.982	0.975	0.968
201 to 400 days	0.989	1	0.982	0.980	0.970
401 to 600 days	0.982	0.982	1	0.962	0.958
601 to 800 days	0.975	0.980	0.962	1	0.966
801 days and more	0.968	0.970	0.958	0.966	1

As can be seen, all the correlations are both positive and extremely strong. The strength of the correlations is perhaps surprising, and the issue of data preparation should be noted here. If a significant proportion of genes in a microarray are either expressed at very low levels, thereby effectively silent, or are expressed at a very high level, throughout all the samples, then this could be a factor for the strong correlations observed here. This is something that can arise to the methods involved in the normalisation of the microarray. The correlations in table 4.2

above show that by simply looking at gene expression in each category, it would be difficult to distinguish the categories, and gain meaningful information about them.

4.4 Glioblastoma GRN Construction and Analysis

For the five categories identified in the previous section, GRNs are constructed from the samples belonging to each group. The WGCNA library of functions for R is used to construct a weighted gene correlation network for each category. Having constructed the weighted networks, only the top 0.5% of edges based on edge weight are retained, with all other edges deleted. There are two reasons for this. The first reason is that keeping only the top 0.5% of edges minimizes the effect of noisy data, making it more likely that biologically meaningful interactions are maintained. Secondly, it is computationally very intensive to work with large networks; an example in point is the calculation of betweenness centrality for each of the original ten networks. Using a 64-bit version of R running on a Windows 7 computer with an Intel Core 2 2.67 GHz processor with 4 GB of RAM, this takes around 30 hours each, in the network with the top 0.5% of edges it takes around 20 seconds.

In order to analyse the networks inferred, the network metrics of weighted betweenness, weighted closeness, weighted degree, degree, and eigenvector centrality, are calculated using the igraph [108] library of functions in R. As well as these node level metrics, a number of network level metrics are calculated; these include assortativity, clustering coefficient, and diameter. The igraph library in R will also be used for the identification of cliques in the networks.

In the previously mentioned study by Upton and Arvanitis, rankings were assigned to each node in the five networks for each of the metrics of weighted degree, weighted betweenness, and closeness centrality, which were then added together, and each gene re-ranked based on the total of these scores. This gave a combined metric ranking score for each gene in each category of GRN. Here, we will present these results in tables; and additionally, will also calculate the individual node rankings for the metrics of betweenness, closeness, weighted degree, degree, and eigenvector centrality.

By scoring the top ranked genes identified by each metric, it is possible to quantitatively compare the performance of the metrics in each category of GRN. This addresses the second objective set out in this work; to investigate and identify appropriate graph theory metrics to analyse and compare the GRNs. The text-mining tool Génie is used with a p-value for abstract selection of 0.01 and FDR 0.01, to generate a list of 299 ranked genes for glioblastoma. This list of genes is included as table A.1 in the appendix. A score is assigned to each gene by subtracting its rank from 300; the top ranking gene of IDH1 is therefore assigned as score of 299, and the 299th ranked gene of ESR2 assigned a score of 1.

4.5 Network Level Metrics

Prior to looking at the individual metrics in the five categories of network, the network level metrics for these five networks will be calculated and compared. Note that weighted assortativity is a slight modification on the previously defined assortativity definition, and simply uses the weighted degree value of the node instead of the degree value. It is possible to use other node values such as weighted betweenness, and weighted closeness for modified assortativity scores.

As well as these network level metrics, the largest unique cliques in each network will be calculated and analysed. Table 4.3 below shows the scores for weighted degree assortativity, degree assortativity, diameter, and clustering in the five networks.

TABLE 4.3 NETWORK LEVEL SCORES ACROSS ALL THE GLIOBLASTOMA GRNS INFERRED USING WGCNA

Metric	**200 or less category**	**201 - 400 category**	**401 - 600 category**	**601 - 800 category**	**800+ category**
Weighted Degree Assortativity	-0.341	-0.335	0.032	-0.307	-0.126
Degree Assortativity	-0.338	-0.320	0.029	-0.294	-0.151
Diameter	1.287	1.107	2.152	1.520	2.167
Network Clustering	0.530	0.563	0.617	0.611	0.568
Largest Clique Size	242 genes	251 genes	249 genes	276 genes	147 genes

From the above network level scores, a number of observations can be made. Firstly, there is no clear correlation between any of the metrics and survival category, suggesting that survival time cannot be associated with any metric behaviour. The diameter might be expected to follow a trend as diameter can be indicative of how easily the genes in the network can communicate with each other, but this is not the case. Whilst the diameter value is highest in the highest survival category, there is no correlation in the other categories. The network clustering does not offer any insights either due to the similar values across the categories, nor does either of the assortativity scores. Perhaps the most important observations to be made are regarding the size and composition of the largest unique cliques in the networks. This is examined in depth in the following section.

4.6 Largest Clique Identification and Analysis

Another analysis of interest on a network wide level is to investigate the largest unique cliques in each network. It is of interest to see both the size of these cliques, and also the genes that comprise them. A large number of common genes in the cliques across all the networks would suggest that the same genes are involved in important cellular processes across the glioblastoma life cycle. As well as identifying common and unique genes in these cliques, it is also of interest to see whether these genes have been previously identified as being either glioblastoma subtype signature genes or are present in the list of ranked glioblastoma genes generated by Génie. Finally, the list of genes that comprise the largest unique clique in each network can be checked for enrichment using the GenesetDB website. Table 4.4 below shows whether genes in each cliques have been identified as glioblastoma subtype signature genes, or glioblastoma candidate genes. Later on in the section, the Venn diagram shows the number of unique genes in each category, and how many genes are common to two or more categories.

TABLE 4.4 GLIOBLASTOMA GENES OF INTEREST IN THE LARGEST UNIQUE CLIQUES

	Proneural Subtype Genes	Mesenchymal Subtype Genes	Neural Subtype Genes	Classical Subtype Genes	Number of Génie genes and score
200 or less category clique	8	0	14	1	1, score of 165
201 – 400 survival days category clique	4	0	15	0	1, score of 178

401- 600 survival days category clique	6	0	15	1	0
601-800 survival days category clique	9	0	19	1	1, score of 165
More than 800 survival days category clique	1	0	10	0	0

Looking at the cliques in the network categories in turn, some interesting results can be seen. Starting with the 200 or less category, 24 previously identified glioblastoma genes of interest are present in the clique of 242 genes; 9.91% of the genes in this category have been previously identified as being of interest for glioblastoma. Of interest is the large number of subtype signature genes identified, 23, out of a total of 242 genes, almost 10%. The top ten gene set enrichment results from GeneSetDB for the genes that comprise this clique are shown in table A.2 of the appendix.

In the 201 - 400 category, 20 previously identified glioblastoma genes of interest are present in the clique of 251 genes; 7.97% of the genes in this category have been previously identified as being of interest for glioblastoma. As with the previous category, the most glioblastoma subtype genes are identified in the neural subtype. In total, there are 19 subtype signature genes present in this category. The top ten gene set enrichment results from GeneSetDB for the genes that comprise this clique are shown in table A.3 of the appendix.

In the 401 - 600 category, 22 previously identified glioblastoma genes of interest are present in the clique of 249 genes; 8.84% of the genes in this category have been previously identified as being of interest for glioblastoma. Once again, there are more signature genes associated with the neural subtype than any other subtype; 6 signature genes associated with the neural subtype are identified in this clique, and one gene associated with the classical subtype. In total, there are 22 subtype signature genes present in this category. The top ten gene set enrichment results from GeneSetDB for the genes that comprise this clique are shown in table A.4 of the appendix.

In the 601 - 800 category, 30 previously identified glioblastoma genes of interest are present in the clique of 276 genes; 10.87% of the genes in this category have been previously identified as being of interest for glioblastoma. This is the largest sized unique cliques of any of the categories. As might be expected due to this, this clique contains more subtype signature genes than any other, 29, and also the highest number of proneural subtype signature genes, 19. The top ten gene set enrichment results from GeneSetDB for the genes that comprise this clique are shown in table A.5 of the appendix.

In the more than 800 category, 11 previously identified glioblastoma genes of interest are present in the clique of 147 genes; 7.48% of the genes in this category have been previously identified as being of interest for glioblastoma. This is a noticeably smaller largest unique clique than for any of the other categories, and also a noticeably smaller amount of previously identified glioblastoma genes of interest present in this clique. It is interesting to note that the smallest sized clique, and the smallest number of previously identified glioblastoma genes of interest are in the GRN for the longest survival category. Once again, there are more signature genes associated with the neural subtype than any other subtype. 10 signature genes associated with the neural subtype are identified in this clique, and one gene associated with the proneural subtype. In total, there are 11 subtype signature genes present in this category. The top ten gene set enrichment results from GeneSetDB for the genes that comprise this clique are shown in table A.6 of the appendix.

The results in table 4.4 suggest quite a high degree of biological significance; a significant amount of genes previously identified with glioblastoma have been identified in all of the cliques. As noted, a noticeably smaller amount of these genes have been identified in the largest unique clique in the highest survival category, 11, compared to 24, 20, 22, and 30 in the other categories, suggesting that certain genes associated with glioblastoma are not as involved in cellular processes for this category as they are in the others. This result could suggest that the previously identified genes are not involved in glioblastoma cellular processes until a later stage of the glioblastoma life cycle. In contrast, a large number of neural subtype signature genes have been identified in the largest cliques in all of the networks. This would indicate that neural signature genes are involved right across the glioblastoma life cycle, and perhaps are not a good indicator of glioblastoma progression due to being ubiquitous throughout the glioblastoma life cycle.

Looking at the distribution of unique and common genes in the cliques, it can be seen that there are 79 genes common to all five cliques, and 144 genes that are unique to one category. Across all five categories there are 1165 genes, of which 431 are unique. This means that only 12.36% of genes appear in one category, and that of the 431 unique genes that are in the cliques, 18.33% appear in all five of the cliques. This suggest that a large number of genes are common to cellular processes across all five categories of the network, and therefore would suggest as a result that these genes are common to cellular processes across the glioblastoma life cycle. There are two ways then to regard these genes; either that they are of interest due to appearing across the glioblastoma life cycle, or to disregard them as it is of interest to identify genes that are unique to specific evolutionary stages. Figure B.1 in the appendix displays the distribution of common and unique genes in the cliques.

5 of these 79 genes have been identified as being neural subtype signature genes although no other subtype signature genes have been identified in this list of 79 genes. It is perhaps not surprising to find a large number of common genes being responsible for cellular processes across the whole glioblastoma life cycle; referring to table 4.2 in a previous section, the very strong positive correlation between the gene expression level rankings across all the five categories can be seen. As such, it could be derived from this that there are a number of common genes involved in cellular processes across all five categories.

4.7 Node Level Metrics

Having looked at the network level metrics, the next step is to focus on the node level metrics. A number of metrics for each node are calculated; weighted degree, degree, weighted betweenness, weighted closeness, and eigenvector centrality. Additionally, a combined metric ranking based on weighted degree, weighted closeness, and weighted betweenness is also calculated, as was the case in the previously highlighted study by Upton and Arvanitis. For each category of GRN, a table of the top 20 ranked genes for each of these metrics is presented, detailing also whether have been previously identified as a glioblastoma subtype signature gene or in the ranked glioblastoma gene list generated by Génie.

The list of ranked glioblastoma genes from the text-mining tool Génie, used previously in this chapter, is used to assign scores to the genes identified by the different metrics. This provides an objective means to score the genes that the different metrics identify. The subjective nature of interpreting the results from GeneSetDB was outlined earlier in the chapter, making it difficult to use these as a basis for scoring the genes identified by the metrics. The enrichment results in this work are used as a guide to the presence of potentially biologically significant results; in order to correctly and meaningfully interpret them, expert biological knowledge is necessary. The scores from the glioblastoma gene list generated by Génie of the top 20, 100, and 500 ranking genes for each metric will be used to compare the performance of the metrics in each category of glioblastoma GRN.

It should be noted that many of the metrics identify the same genes in the GRNs; this can also be seen in the high Spearman correlation scores between the metrics. Due to this repetition in the genes identified by the different metrics, for the basis of comparing the top ranking genes identified in each category, the combined metric ranking based on the metrics of weighted betweenness, weighted closeness, and weighted degree centrality will be used.

4.7.1 200 or fewer Survival Days Category

Table 4.5 below shows the top 20 ranked genes identified by each of the metrics. Additionally, the table also details whether the gene has previously been identified as either a glioblastoma subtype signature gene or is present of the ranked list of glioblastoma genes generated by the text-mining tool Génie.

TABLE 4.5 TOP 20 RANKED GENES IDENTIFIED BY EACH METRIC IN 0 – 200 SURVIVAL DAYS CATEGORY GRN

Degree Rank	Gene	Previously Identified	Weighted Degree Rank	Gene	Previously Identified	Weighted Betweenness Rank	Gene	Previously Identified
1	SYN1	No	1	SYN1	No	1	RAB40B	No
2	AK5	No	2	SULT4A1	No	2	DNAJC6	No
3	SULT4A1	No	3	AK5	No	3	ATP8A1	No
4	SLC17A7	No	4	SLC17A7	No	4	MAST3	No
5	HSPA12A	No	5	PAK6	No	5	TUBB4	No

6	PAK6	No	6	HSPA12A	No	6	ATP2B2	No
7	SH3GL2	Yes, c	7	KIAA0513	No	7	ANK3	No
8	NAP1L2	No	8	STXBP1	No	8	PAK3	Yes, a
9	STXBP1	No	9	SH3GL2	Yes, c	9	NCF4	Yes, b
10	GRM5	No	10	NAP1L2	No	10	SLC25A4	No
11	KIAA0513	No	11	GRM5	No	11	SEC14L5	No
12	SV2B	No	12	SV2B	No	12	KIF5A	No
13	MOAP1	No	13	GLS2	No	13	SYT1	No
14	GLS2	No	14	SLC12A5	No	14	PRKCZ	No
15	SCN2A	No	15	PHYHIP	No	15	NEUROD2	No
16	CDH18	No	16	SNAP91	Yes, a	16	GNAO1	No
17	SNAP91	Yes, a	17	SCN2A	No	17	MAP1A	No
18	SLC12A5	No	18	CDH18	No	18	SCN2A	No
19	EPB41L1	No	19	MOAP1	No	19	AATK	No
20	NSF	No	20	EPB41L1	No	20	C1orf38	Yes, b
Weighted Closeness Rank	**Gene**	**Previously Identified**	**Eigenvector Centrality Rank**	**Gene**	**Previously Identified**	**Combined Metric Rank**	**Gene**	**Previously Identified**
1	AK5	No	1	SYN1	No	1	AK5	No
2	HSPA12A	No	2	SULT4A1	No	2	SYT1	No
3	SYT1	No	3	SLC17A7	No	3	MAP1A	No
4	SYN1	No	4	PAK6	No	4	HSPA12A	No
5	MAP1A	No	5	KIAA0513	No	5	SNAP91	Yes, a
6	SLC17A7	No	6	AK5	No	6	SULT4A1	No
7	STXBP1	No	7	HSPA12A	No	7	EPB41L1	No
8	PRKCZ	No	8	STXBP1	No	8	SCN2A	No
9	EPB41L1	No	9	GLS2	No	9	PRKCZ	No
10	MAST3	No	10	NAP1L2	No	10	STXBP1	No
11	SNAP91	Yes, a	11	SH3GL2	Yes, c	11	SLC17A7	No
12	NRGN	No	12	PHYHIP	No	12	MAST3	No
13	SULT4A1	No	13	SLC12A5	No	13	SYN1	No
14	DYNC1I1	Yes, c	14	GRM5	No	14	PAK6	No
15	RGS7	No	15	SV2B	No	15	VAMP2	No

16	ANK3	No	16	SNAP91	Yes, a	16	RGS7	No
17	VAMP2	No	17	CDH18	No	17	MOAP1	No
18	RAB6B	No	18	NSF	No	18	NAP1L2	No
19	KIF3C	No	19	SCN2A	No	19	NRGN	No
20	CYFIP2	No	20	OLFM1	No	20	RAB6B	No

a = proneural signature gene, b = Mesenchymal signature gene, c = neural signature gene, d = classical signature gene

A number of glioblastoma subtype signature genes are identified by the metrics, as can be seen in table 4.5 above. Note that a number of common genes are identified by the metrics, as shown by the Spearman rank correlation scores for the genes identified by the metrics, shown in table 4.6 below, and the Venn diagram of common and unique genes identified by the metrics, figure B.2 of the appendix.

TABLE 4.6 SPEARMAN RANK CORRELATION SCORES OF RANKINGS ASSIGNED BY THE METRICS TO THE GENES IN 0 – 200 SURVIVAL DAYS CATEGORY GRN

	Degree Centrality	**Weighted Degree**	**Weighted Betweenness**	**Weighted Closeness**	**Eigenvector Centrality**
Degree Centrality	1	0.998	0.534	0.740	0.764
Weighted Degree	0.998	1	0.537	0.726	0.755
Weighted Betweenness	0.534	0.537	1	0.177	0.199
Weighted Closeness	0.740	0.726	0.177	1	0.892
Eigenvector Centrality	0.764	0.755	0.199	0.892	1

All of the metrics apart from weighted betweenness centrality correlate strongly, greater than +0.7, and identify similar genes, something that is borne out by only 48 unique genes being identified out of 100. The extremely strong correlation between degree and weighted degree is expected due to the method of only keeping the top 0.5% of edges, and also the strong correlations between degree, weighted degree, and eigenvector centrality. There is also a strong correlation between weighted closeness and eigenvector centrality, implying that they identify the same genes as being of high rank.

If the betweenness centrality metric is removed, only 34 unique genes are identified out of 80 genes. This shows that the betweenness centrality identified 14 genes that the other metrics did not. It should be noted as well that 7 genes were identified by four of the network metrics; all but weighted betweenness centrality. A visual representation of this is shown in the Venn diagram in figure B.2 of the appendix, showing 14 unique genes are identified using weighted betweenness, 7 using weighted closeness, and 2 using eigenvector centrality.

For the purposes of this chapter, and the work as a whole, presenting the genes identified as table 4.5 does not help to address the objectives. It does not provide an objective means by which to score the metrics, as scoring a gene based on whether it is a glioblastoma subtype signature genes adds subjectivity requiring specialist biological knowledge. Additionally, a table with a large number of genes detailed may provide the reader with a lot of unnecessary information. As such, the same approach adopted in our previous study will be used; the top 20 ranked genes for the combined metric will solely be presented for each category of GRN for the rest of this chapter, and the rest of the thesis.

One of the main reasons for using various metrics was to investigate whether certain metrics performed better than others in giving biologically significant results. Therefore, a better approach is to score the top 20, top 100, and top 500 ranked genes for each metric using the list generated by Génie. This directly addresses the second objective; it gives a means by which to score the metrics for the genes that they identify in the GRNs for the different categories of glioblastoma, allowing the performance of the metric in each category to be compared based on their ability to identify genes in the Génie list. Starting with the top 20 ranked genes by each metric, table 4.7 below shows the scores that are obtained.

TABLE 4.7 GÉNIE SCORES AND METRIC RANKING FOR THE TOP 20 RANKED GENES FOR EACH METRIC IN 0 – 200 SURVIVAL DAYS CATEGORY GRN

Metric	**Number of Génie glioblastoma genes identified**	**Génie Score**	**Metric Rank**
Degree Centrality	0	0	1st
Weighted Degree	0	0	1st
Weighted Betweenness	0	0	1st
Weighted Closeness	0	0	1st
Eigenvector Centrality	0	0	1st
Combined Metric	0	0	1st

None of the metrics identify a single Génie glioblastoma gene of interest. Applying the scoring system to the top 100 ranked genes for each metric, the results in table 4.8 below are given.

TABLE 4.8 GÉNIE SCORES AND METRIC RANKING FOR THE TOP 100 RANKED GENES FOR EACH METRIC IN 0 – 200 SURVIVAL DAYS CATEGORY GRN

Metric	**Number of Génie glioblastoma genes identified**	**Génie Score**	**Metric Rank**
Degree Centrality	1	165	1st
Weighted Degree	1	165	1st

Weighted Betweenness	1	165	1st
Weighted Closeness	1	165	1st
Eigenvector Centrality	1	165	1st
Combined Metric	1	165	1st

All of the metrics identify the same glioblastoma gene of interest, LGI1, ranked as the 135th most important gene for glioblastoma. Finally, extending the scoring system to the top 500 genes, the following results shown in table 4.9 below are given.

TABLE 4.9 GÉNIE SCORES AND METRIC RANKING FOR THE TOP 500 RANKED GENES FOR EACH METRIC IN 0 – 200 SURVIVAL DAYS CATEGORY GRN

Metric	**Number of Génie glioblastoma genes identified**	**Génie Score**	**Metric Rank**
Weighted Betweenness	9	1107	1st
Eigenvector Centrality	5	364	2nd
Weighted Closeness	3	295	3rd
Degree Centrality	5	281	4th
Weighted Degree	5	281	4th
Combined Metric	4	238	6th

The weighted betweenness is the best performing metric, followed by the eigenvector centrality. The degree and weighted degree centrality both identify the same 5 genes, and the weighted closeness centrality identifies 3 genes. The combined metric ranking performs the worst; although it identifies one more gene than the weighted closeness, the genes it identifies are lower scoring. Only one of the 9 genes identified by the weighted betweenness centrality is in the top 50 Génie ranked genes for glioblastoma. In total, 33 out of the 299 Génie glioblastoma ranked genes are present in this network.

These results might suggest that using the combined metric ranking will not lead to the identification of biologically significant genes. It should be taken into account that this is solely for one network, and that the different metrics will perform differently across the different networks. The main purpose of using different metrics is to objectively score them based on their ability to identify genes from a ranked list; it is of interest to identify high ranking genes for the metrics that have been identified as glioblastoma subtype signature genes or have been identified in previous glioblastoma studies, but it does not specifically address the objectives laid out in the first chapter. Furthermore, presenting only one list of 20 genes for each category allows the interrogation to be much more thorough, and in addition to the use of GenseSetDB, The Cancer Genome Atlas [109] , the IntOGen browser [110], and CGPrio [111], will be used to aid in the biological interpretation of the results returned for the glioblastoma GRNs.

The enrichment results for the top 20 ranked genes for the combined metric are shown in table A.7 of the appendix. Nothing of note for glioblastoma stands out from these results. The most notable gene in the top 20 combined metric ranking for this category is SNAP91, which has been identified in a number of studies as a signature gene of the proneural glioblastoma subtype, such as the study by Brennan et al [112]. This is the only signature gene of any glioblastoma subtype that occurs in the top 20 ranking. The 2nd ranked gene in this list, SYT1, is highly ranked as an oncogene by CGPrio, and is also identified in a glioblastoma study by Dong et al [113] as being a candidate gene for the disease. The 9th ranked gene, PRKCZ, has been identified in 3 glioblastoma gene lists by the cancer genome atlas gene checker. PRKCZ was also been shown in a study by Donson et al [114] to be crucial to proliferation in glioblastoma cell lines. 9 of the top 20 ranked genes in this category do not appear in the top 20 rankings for any of the other categories, including the top 4 ranked genes. This suggests that these genes identified as being important in this category do not have such an important role in the other categories of network.

Of the 11 genes in the top 20 combined metric list that appear in top 20 combined metric lists for other categories; 3 genes, SNAP91, SYN1, and RGS7, appear in three of the five top 20 lists, including this category. SNAP91 has already been highlighted; however the two other genes have not previously been identified as candidate glioblastoma genes, and this suggests that they may play a role across various stages of the glioblastoma life cycle. The other two genes, SYN1 and RGS7, also appear in all five of the largest unique cliques across the different categories, which would also suggest that they are involved in cellular processes across the whole glioblastoma life cycle. The top ten enrichment results for this list of genes are shown in table XVII. Despite a number of enrichment results yielded, none of the results in the top ten relate to glioblastoma.

4.7.2 201-400 Survival Days Category

Table 4.10 on the following page shows the top 20 ranked genes for the combined metric in the 201-400 survival days category GRN. Looking at the top 20 ranked genes for the combined metric, the presence of SNAP91 amongst the top 20 ranked genes once again stands out. The second result of note is that the two top ranked genes for the combined metric are both solute carriers; SLC9A6 is a sodium/hydrogen exchanger, and SLC8A2 is involved in sodium/calcium exchange. This result would suggest that the exchange of sodium plays a role in glioblastomas within the 201-400 survival days category, and potentially is an area of interest for glioblastoma studies. 10 of the top 20 ranked genes for the combined metric in this category do not appear in the top 20 combined metric rankings for any of the other categories, including the top 2 ranked genes previously discussed. This potentially suggests that the importance of sodium exchange is limited to glioblastomas within this category, and that, as before, there are a number of genes identified as being important in this category that do not have such an important role in other categories.

Of the 10 genes that do appear in top 20 combined metric ranked lists in other categories, there are again 3 genes that appear in three of the five top 20 lists, including this category. As well as SNAP91, these include PPP1R16B, and CAMKV. Whilst these two genes have not been specifically identified as candidate glioblastoma genes, it should be noted that CAMKV has been identified as potential cancer gene target by the Broad Institute research group [115]. These two genes, PPP1R16B and CAMKV, are also part of the list of genes that appear in all of the largest cliques across all of the categories in the network, suggesting that they are involved in cellular processes across the glioblastoma life cycle.

The top 10 results for this list of genes from GenesetDB are shown in table A.8 in the appendix. The 8th result, relating to ERBB1 pathway, should be noted; ERBB1 has been identified as being over expressed in high grade glioma, in a 2003 study by Gilbertson et al [116].

TABLE 4.10 TOP 20 RANKED GENES FOR COMBINED METRIC IN 201 – 400 SURVIVAL DAYS CATEGORY GRN

Gene	***Combined Metric Rank***	***Gene***	***Combined Metric Rank***
SLC9A6	1	CAMKV	11
SLC8A2	2	DYNC1I1	12

INA	3	PHYHIP	13
SNAP91	4	KCNAB2	14
PPP1R16B	5	NAP1L2	15
MEF2C	6	MAST3	16
WDR7	7	TAGLN3	17
CYFIP2	8	FBXO41	18
KIAA0513	9	STXBP1	19
PDE2A	10	MOAP1	20

Table 4.11 below shows the Spearman Rank correlation scores for the rankings assigned to the genes by each metric in the 201-400 survival days category GRN inferred using WGCNA.

TABLE 4.11 SPEARMAN RANK CORRELATION SCORES OF RANKINGS ASSIGNED BY THE METRICS TO THE GENES IN THE 201-400 SURVIVAL DAYS CATEGORY GRN

	Degree Centrality	**Weighted Degree**	**Eigenvector Centrality**	**Closeness Centrality**	**Betweenness Centrality**
Degree Centrality	1	0.998	0.790	0.831	0.664
Weighted Degree	0.998	1	0.772	0.814	0.673
Eigenvector Centrality	0.790	0.772	1	0.962	0.374
Closeness Centrality	0.831	0.814	0.962	1	0.395
Betweenness Centrality	0.664	0.673	0.374	0.395	1

As was the case with the previous category, all of the metrics apart from weighted betweenness centrality correlate strongly, > +0.75, and as such identify similar genes. The correlations are even stronger this time compared to the last category, and the weighted betweenness has a stronger correlation with the other metrics as well compared to the last category. Using the scoring system for the metrics used in the previous category, the following rankings for the metrics based on their top 20 ranked genes are shown in table 4.12 below.

TABLE 4.12 GÉNIE SCORES AND METRIC RANKING FOR THE TOP 20 RANKED GENES FOR EACH METRIC IN 201 – 400 SURVIVAL DAYS CATEGORY GRN

Metric	**Number of Génie glioblastoma genes identified**	**Génie Score**	**Metric Rank**
Degree Centrality	0	0	1st
Weighted Degree	0	0	1st
Weighted Betweenness	0	0	1st
Weighted Closeness	0	0	1st
Eigenvector Centrality	0	0	1st
Combined Metric	0	0	1st

None of the metrics identify a single Génie glioblastoma gene of interest. Applying the scoring system to the top 100 ranked genes for each metric, the results in the table 4.13 below are given.

TABLE 4.13 GÉNIE SCORES AND METRIC RANKING FOR THE TOP 100 RANKED GENES FOR EACH METRIC IN 201 – 400 SURVIVAL DAYS CATEGORY GRN

Metric	**Number of Génie glioblastoma genes identified**	**Génie Score**	**Metric Rank**
Degree Centrality	0	0	1st
Weighted Degree	0	0	1st
Weighted Betweenness	0	0	1st
Weighted Closeness	0	0	1st
Eigenvector Centrality	0	0	1st
Combined Metric	0	0	1st

Again, none of the metrics identify a single Génie glioblastoma gene of interest. Despite the enrichment result pertaining to ERBB1, the metrics do not identify any glioblastoma genes of interest. Applying the scoring system to the top 500 ranked genes for each metric, the results in table 4.14 below are given.

TABLE 4.14 GÉNIE SCORES AND METRIC RANKING FOR THE TOP 500 RANKED GENES FOR EACH METRIC IN 201 – 400 SURVIVAL DAYS CATEGORY GRN

Metric	**Number of Génie glioblastoma genes identified**	**Génie Score**	**Metric Rank**
Weighted Betweenness	6	944	1st
Eigenvector Centrality	5	499	2nd
Weighted Closeness	4	412	3rd
Degree Centrality	5	370	4th
Weighted Degree	5	370	4th

Combined Metric	5	370	4th

The weighted betweenness is the best performing metric, followed by the eigenvector centrality, as was the case with the last glioblastoma category. The degree, weighted degree and combined metric each identify the same 5 genes. Again, only one of the 9 genes identified by the weighted betweenness centrality is in the top 50 Génie ranked genes for glioblastoma, the gene TERT. In total, 33 out of the 299 Génie glioblastoma ranked genes are present in this network. It should be noted that this category and the previous category have the same number of Génie glioblastoma ranked genes in their networks, although they are not the same genes.

4.7.3 401-600 Survival Days Category

Table 4.15 below shows the top 20 ranked genes for the combined metric in the 401-600 survival days category GRN. The previously mentioned genes CAMKV, SYN1 and RGS7 appear in this list, and are again the only genes that appear in two other top 20 ranked lists, as well as this one. The presence of gene EPHB6 is interesting; as well as being identified in one glioblastoma specific gene list, it has also been identified as being on six other cancer gene lists, such as ovarian cancer and breast cancer, by the cancer genome atlas. SH3GL2 is another glioblastoma gene of interest identified in the glioblastoma study by Dong et al, and also a 2008 study by Chang [117]. It is also a signature gene of the neural glioblastoma subtype, as are the genes CPNE6, and HPCAL4. It is worth noting the presence of 3 neural subtype signature genes in the top 20 ranked genes for this category.

13 of the 20 genes that appear in this list do not appear on any of the other top 20 genes lists. This is a greater number of unique genes than the previous two lists, suggesting increasingly

different network behaviour as the survival time increases. It also suggests that genes that are important for the behaviour of the network in the two previous categories are not as important for the function of the network in this category, and that different processes are taking place that these genes are not involved in. None of the top 10 results for this list of genes from GenesetDB, shown in table A.9 in the appendix stand out as being of interest to glioblastoma.

TABLE 4.15 TOP 20 RANKED GENES FOR COMBINED METRIC IN 401 – 600 SURVIVAL DAYS CATEGORY GRN

Gene	***Combined Metric Rank***	***Gene***	***Combined Metric Rank***
RIMS3	1	BZRAP1	11
CAMKV	2	SH3GL2	12
CA11	3	PCSK2	13
CHGB	4	PRKAR1B	14
GLS2	5	HPCAL4	15
NELL2	6	MYRIP	16
EPHB6	7	CPNE6	17
INA	8	SYN1	18
GAD2	9	PAK6	19
KIAA1107	10	RGS7	20

Table 4.16 below shows the Spearman Rank correlation scores for the rankings assigned to the genes by each metric in the 401-600 survival days category GRN inferred using WGCNA.

TABLE 4.16 SPEARMAN RANK CORRELATION SCORES OF RANKINGS ASSIGNED BY THE METRICS TO THE GENES IN THE 401-600 SURVIVAL CATEGORY GRN

	Degree Centrality	**Weighted Degree**	**Eigenvector Centrality**	**Closeness Centrality**	**Betweenness Centrality**
Degree Centrality	1	0.998	0.704	0.698	0.715
Weighted Degree	0.998	1	0.695	0.684	0.705
Eigenvector Centrality	0.704	0.695	1	0.904	0.416
Closeness Centrality	0.698	0.684	0.904	1	0.494
Betweenness Centrality	0.715	0.705	0.416	0.494	1

Compared to the last category, the correlations are weaker between all of the metrics. As before, weighted closeness and eigenvector centrality have a very strong correlation, >0.9, but the other correlations are noticeably weaker. Whilst before the correlations between all of the metrics apart from weighted betweenness was at least >0.75, this time they are >0.68. This implies that for this category there is a greater diversity in the genes that the metrics identify. Applying the scoring metrics used in the previous category to the top 20 ranked genes for each metric, the rankings shown in table 4.17 are obtained.

TABLE 4.17 GÉNIE SCORES AND METRIC RANKING FOR THE TOP 20 RANKED GENES FOR EACH METRIC IN 401 – 600 SURVIVAL DAYS CATEGORY GRN

Metric	**Number of Génie glioblastoma genes identified**	**Génie Score**	**Metric Rank**
Weighted Betweenness	1	273	1st
Degree Centrality	0	0	2nd
Weighted Degree	0	0	2nd
Weighted Closeness	0	0	2nd
Eigenvector Centrality	0	0	2nd
Combined Metric	0	0	2nd

Weighted betweenness is the only metric to identify a glioblastoma gene of interest, the 27th ranked PLAUR. Applying the scoring system to the top 100 ranked genes for each metric, the results in table 4.18 below are given.

TABLE 4.18 GÉNIE SCORES AND METRIC RANKING FOR THE TOP 100 RANKED GENES FOR EACH METRIC IN 401 – 600 SURVIVAL DAYS CATEGORY GRN

Metric	**Number of Génie glioblastoma genes identified**	**Génie Score**	**Metric Rank**
Weighted Betweenness	3	610	1st
Degree Centrality	0	0	2nd
Weighted Degree	0	0	2nd
Weighted Closeness	0	0	2nd
Eigenvector Centrality	0	0	2nd
Combined Metric	0	0	2nd

Weighted betweenness centrality is again the only metric that identifies any glioblastoma genes of interest, and in addition this time also identifies the genes RAC2, ranked 164th most important gene for glioblastoma, and EZH2, ranked 99th most important gene for glioblastoma. Finally, applying the scoring system to the top 500 ranked genes for each metric, the results in table 4.19 below are given.

TABLE 4.19 GÉNIE SCORES AND METRIC RANKING FOR THE TOP 500 RANKED GENES FOR EACH METRIC IN 401 – 600 SURVIVAL DAYS CATEGORY GRN

Metric	**Number of Génie glioblastoma genes identified**	**Génie Score**	**Metric Rank**
Weighted Betweenness	10	1895	1st
Weighted Degree	5	830	2nd
Degree Centrality	4	617	3rd
Weighted Closeness	4	302	4th
Eigenvector Centrality	4	302	4th
Combined Metric	3	215	6th

The weighted betweenness is the best performing metric, followed by the weighted degree, and degree centrality. The combined metric is the worst performing metric, identifying 3 genes. Of the 10 genes identified by the weighted betweenness centrality, 3 are in the top 50 Génie ranked genes for glioblastoma. This suggests that the weighted betweenness centrality performs better at identifying glioblastoma genes of interest in this glioblastoma category than the previous two. In total, 44 out of the 299 Génie glioblastoma ranked genes are present in this network, 11 more than the previous two categories. This might seem surprising; it might be expected that the networks inferred for the previous two glioblastoma stages would contain more glioblastoma genes due to their shorter survival time, however Génie does not distinguish the genes based on glioblastoma evolution.

4.7.4 601-800 Survival Days Category

Table 4.20 below shows the top 20 ranked genes for the combined metric in the 601-800 survival days category GRN. The 2nd ranked gene, VAMP2, which is the 15th ranked gene in the lowest survival category network, is ranked by IntOGen as having a high probability of being an oncogene, as is the 5th ranked gene SCAMP5. This would suggest that these genes are highly likely to be involved in interactions with other genes, and the high ranking results here concur with that prediction. The gene PRKCZ is the 20th ranked gene in this list, having previously been highlighted as being fundamental in glioblastoma proliferation in human cell lines.

There are again 13 unique entries on the top 20 ranking list for this category of network, suggesting, as was the case with the last network category, that there is markedly different behaviour in the network, compared to the other categories of network. The three genes in this list that also occur in two other lists are PPP1R16B, RGS7, and the proneural signature gene SNAP91. Once again, despite a number of enrichment results yielded for this list of genes, shown in table A.10 of the appendix, none of the top ten results specifically relate to glioblastoma.

TABLE 4.20 TOP 20 RANKED GENES FOR COMBINED METRIC IN 601 – 800 SURVIVAL DAYS CATEGORY GRN

Gene	*Combined Metric Rank*	*Gene*	*Combined Metric Rank*
TUBB	1	SNAP91	11
VAMP2	2	ATP6V1G2	12
HLF	3	RGS7	13
ARHGEF9	4	GFOD1	14
SCAMP5	5	PDE2A	15
PPP3CB	6	ATP2B2	16
SNPH	7	CHGA	17
PPP1R16B	8	EPB49	18
GOT1	9	S100A1	19
IQSEC3	10	PRKCZ	20

Table 4.21 below shows the Spearman Rank correlation scores for the rankings assigned to the genes by each metric in the 601-800 survival days category GRN inferred using WGCNA.

TABLE 4.21 SPEARMAN RANK CORRELATION SCORES OF RANKINGS ASSIGNED BY THE METRICS TO THE GENES IN THE 601-800 SURVIVAL CATEGORY GRN

	Degree Centrality	**Weighted Degree**	**Eigenvector Centrality**	**Closeness Centrality**	**Betweenness Centrality**
Degree Centrality	1	0.999	0.824	0.864	0.754
Weighted Degree	0.999	1	0.811	0.852	0.757

Eigenvector Centrality	0.824	0.811	1	0.976	0.524
Closeness Centrality	0.864	0.852	0.976	1	0.565
Betweenness Centrality	0.754	0.757	0.524	0.565	1

For this category, as was the case with 201-400 category, all of the metrics apart from weighted betweenness centrality correlate strongly, > +0.80, and as such identify similar genes. The correlations are even stronger than for the 201-400 category, and the weighted betweenness has a stronger correlation with the other metrics as well. Note the extremely strong correlations between eigenvector centrality and closeness centrality, >0.97.

Applying the Génie scoring system to the top 20 ranked genes for each metric, the following metric rankings shown in table 4.22 below are obtained.

TABLE 4.22 GÉNIE SCORES AND METRIC RANKING FOR THE TOP 20 RANKED GENES FOR EACH METRIC IN 601 – 800 SURVIVAL DAYS CATEGORY GRN

Metric	Number of Génie glioblastoma genes identified	Génie Score	Metric Rank
Weighted Betweenness	2	268	1st
Degree Centrality	0	0	2nd
Weighted Degree	0	0	2nd
Weighted Closeness	0	0	2nd
Eigenvector Centrality	0	0	2nd
Combined Metric	0	0	2nd

All of the metrics apart from the weighted betweenness centrality fail to identify any ranked glioblastoma genes of interest. Weighted betweenness centrality identifies two glioblastoma genes of interest; FPR1 ranked 168th for glioblastoma, and RAC2 ranked 164th for glioblastoma. Applying the scoring system to the top 100 ranked genes for each metric, the results in table 4.23 below are given.

TABLE 4.23 GÉNIE SCORES AND METRIC RANKING FOR THE TOP 100 RANKED GENES FOR EACH METRIC IN 601 – 800 SURVIVAL DAYS CATEGORY GRN

Metric	**Number of Génie glioblastoma genes identified**	**Génie Score**	**Metric Rank**
Weighted Betweenness	3	488	1st
Degree Centrality	0	0	2nd
Weighted Degree	0	0	2nd
Weighted Closeness	0	0	2nd
Eigenvector Centrality	0	0	2nd
Combined Metric	0	0	2nd

Weighted betweenness centrality again is the only metric that identifies any glioblastoma genes of interest; this time, in addition to the genes FPR1 and RAC2, the 80th ranked gene SOX10 is identified. Applying the scoring system to the top 500 ranked genes for each metric, the results in table 4.24 below are given.

TABLE 4.24 GÉNIE SCORES AND METRIC RANKING FOR THE TOP 500 RANKED GENES FOR EACH METRIC IN 601 – 800 SURVIVAL DAYS CATEGORY GRN

Metric	**Number of Génie glioblastoma genes identified**	**Génie Score**	**Metric Rank**
Weighted Betweenness	6	1036	1st
Weighted Closeness	3	575	2nd
Combined Metric	3	563	3rd
Weighted Degree	4	397	4th
Degree Centrality	2	343	5th
Eigenvector Centrality	2	343	5th

The weighted betweenness is the best performing metric, followed by the weighted closeness, and the combined metric. Degree and eigenvector centrality are the worst performing metrics, identifying 2 genes. None of the 6 genes identified by the weighted betweenness centrality are in the top 50 Génie ranked genes for glioblastoma. In total, 37 out of the 299 Génie glioblastoma ranked genes are present in this network, 7 fewer than the previous category, but 3 more than the first two categories.

4.7.5 More than 800 Survival Days Category

Table 4.25 below shows the top 20 ranked genes in the 801+ survival days category. The presence of GABRD and GABRA1 as the top two ranked genes is immediately noticeable, suggesting that the GABR area is of interest. In fact, whilst these two genes are not signature genes of any glioblastoma subtype, the gene GABR2 is a proneural signature gene, as identified by Verhaak et al [106]. The 9th ranked gene, KALRN, appears in 3 glioblastoma gene lists in the cancer genome atlas gene ranker, as well as appearing in 3 other cancer gene related lists. The enrichment results

As with the two previous categories, there are 13 unique entries on the top 20 ranking list for this network. There are also 3 genes in the list below that also appear on two other lists, these genes are CAMKV, SYN1, and PPP1R16B. The enrichment results for this list of genes are shown in table A.11 of the appendix, however none of the top ten results specifically relate to glioblastoma.

TABLE 4.25 TOP 20 RANKED GENES FOR COMBINED METRIC IN MORE THAN 800 SURVIVAL DAYS CATEGORY GRN

Gene	***Combined Metric Rank***	***Gene***	***Combined Metric Rank***
GABRD	1	GLS2	11
GABRA1	2	CACNG3	12
EPB41L1	3	SLC12A5	13
BSN	4	KIAA1107	14
CAMKV	5	PNOC	15
CALY	6	NUAK1	16
FAM153A	7	SYN2	17
SYN1	8	CABP1	18
KALRN	9	GOT1	19
PPP1R16B	10	SNCB	20

Table 4.26 below shows the Spearman Rank correlation scores for the rankings assigned to the genes by each metric in the more than 800 survival days category GRN inferred using WGCNA.

TABLE 4.26 SPEARMAN RANK CORRELATION SCORES OF RANKINGS ASSIGNED BY THE DIFFERENT METRICS TO THE GENES IN MORE THAN 800 SURVIVAL CATEGORY GRN

	Degree Centrality	**Weighted Degree**	**Eigenvector Centrality**	**Closeness Centrality**	**Betweenness Centrality**
Degree Centrality	1	0.995	0.692	0.766	0.703
Weighted Degree	0.995	1	0.676	0.744	0.695
Eigenvector Centrality	0.692	0.676	1	0.638	0.445
Closeness Centrality	0.766	0.744	0.638	1	0.647
Betweenness Centrality	0.703	0.695	0.445	0.647	1

The correlations between the metrics are not as strong as the last category. Only the correlation between degree and weighted degree is stronger than 0.8 this time. However, the correlations between the weighted betweenness and the other metrics are stronger this time compared to the previous categories, with only the correlation between weighted betweenness and eigenvector centrality less than 0.64. This implies that the weighted betweenness will identify more genes in common with the other metrics, compared to the other categories. Using the scoring system for the metrics, the following rankings shown in table 4.27 based on the top 20 ranked genes for each metric are obtained.

TABLE 4.27 GÉNIE SCORES AND METRIC RANKING FOR THE TOP 20 RANKED GENES FOR EACH METRIC IN MORE THAN 800 SURVIVAL DAYS CATEGORY GRN

Metric	Number of Génie glioblastoma genes identified	Génie Score	Metric Rank
Degree Centrality	0	0	1st
Weighted Degree	0	0	1st
Weighted Betweenness	0	0	1st
Weighted Closeness	0	0	1st
Eigenvector Centrality	0	0	1st
Combined Metric	0	0	1st

None of the metrics identify a single Génie glioblastoma gene of interest. Applying the scoring system to the top 100 ranked genes for each metric, the results in table 4.28 below are given.

TABLE 4.28 GÉNIE SCORES AND METRIC RANKING FOR THE TOP 100 RANKED GENES FOR EACH METRIC IN MORE THAN 800 SURVIVAL DAYS CATEGORY GRN

Metric	Number of Génie glioblastoma genes identified	Génie Score	Metric Rank
Weighted Betweenness	1	68	1st
Degree Centrality	2	54	2nd
Weighted Degree	2	54	2nd
Weighted Closeness	0	0	4th
Eigenvector Centrality	0	0	4th
Combined Metric	0	0	4th

Weighted betweenness is the best performing metric, identifying one ranked glioblastoma gene, the 232nd ranked ALOX2. Weighted degree and degree centrality are the next best performing metrics, both identifying the same two genes; ING4 ranked 296th most important gene for glioblastoma, and CSF2 ranked 250th most important gene for glioblastoma. The other metrics do not identify any ranked glioblastoma genes. Applying the scoring system to the top 500 ranked genes for each metric, the results in the table below are given.

TABLE 4.29 GÉNIE SCORES AND METRIC RANKING FOR THE TOP 500 RANKED GENES FOR EACH METRIC IN MORE THAN 800 SURVIVAL DAYS CATEGORY GRN

Metric	**Number of Génie glioblastoma genes identified**	**Génie Score**	**Metric Rank**
Weighted Betweenness	10	804	1st
Weighted Closeness	6	626	2nd
Eigenvector Centrality	6	626	2nd
Degree Centrality	5	522	4th
Weighted Degree	5	522	4th
Combined Metric	5	450	6th

The weighted betweenness is the best performing metric, followed by the weighted closeness eigenvector centrality. The combined metric is the worst performing metric; despite identifying the same number of genes as degree and weighted degree, the genes identified are lower scoring. None of the 10 genes identified by the weighted betweenness centrality, the best performing metric, are in the top 50 Génie ranked genes for glioblastoma. In total, 34 out of the 299 Génie glioblastoma ranked genes are present in this network, the same number as in the first two categories.

4.8 Overall Metric Scores

Previously, each of the metrics was ranked based on the top 20, top 100 and top 500 genes they identified for each glioblastoma GRN category. This ranking is based on the score of the identified genes in the Génie glioblastoma gene list.

The results of these metric rankings can now be presented together, allowing an analysis of which metric is the best performing overall. Assigning an equal weighting to the scores of the top 20, top 100, and top 500 ranked genes for each metric in each category, the following ranks shown in table 4.30 below are given.

TABLE 4.30 OVERALL METRIC RANKS IN THE DIFFERENT CATEGORIES OF GLIOBLASTOMA GRN INFERRED USING WGCNA

	200 or fewer category rank	201-400 category rank	401-600 category rank	601-800 category rank	More than 800 category rank	Ranking totals	Overall rank
Weighted Betweenness	1st	1st	1st	1st	1st	5	1st
Weighted Closeness	3rd	3rd	4th	2nd	3rd	15	2nd
Eigenvector Centrality	2nd	2nd	4th	5th	3rd	16	3rd
Weighted Degree	4th	4th	2nd	4th	3rd	17	4th
Degree Centrality	4th	4th	3rd	5th	3rd	19	5th
Combined Metric	6th	4th	6th	3rd	6th	25	6th

From table 4.30 above it can be seen that weighted betweenness is clearly the best performing metric overall, followed by weighted closeness and eigenvector centrality. This is not surprising, considering that weighted betweenness was the only metric to identify glioblastoma genes of interest a number of times.

Taking just the scores for the top 20 genes identified by each metric in each category, the following rankings shown in table 4.31 are obtained.

TABLE 4.31 METRIC RANKS IN THE DIFFERENT CATEGORIES OF GLIOBLASTOMA GRN INFERRED USING WGCNA BASED ON TOP 20 GENES IDENTIFIED BY EACH METRIC

	200 or fewer category rank	201-400 category rank	401-600 category rank	601-800 category rank	More than 800 category rank	Ranking totals	Overall rank
Weighted Betweenness	1st	1st	1st	1st	1st	5	1st
Degree Centrality	1st	1st	2nd	2nd	1st	7	2nd
Weighted Degree	1st	1st	2nd	2nd	1st	7	2nd
Weighted Closeness	1st	1st	2nd	2nd	1st	7	2nd
Eigenvector Centrality	1st	1st	2nd	2nd	1st	7	2nd
Combined Metric	1st	1st	2nd	2nd	1st	7	2nd

Whilst the weighted betweenness is the best performing metric, followed by the other metrics, it should be noted though that solely using the top 20 genes identified by each metric is not very informative. In a number of the categories, no glioblastoma genes were identified by any of the metrics, giving all the metrics the same ranking. Taking just the scores for the top 100 genes identified by each metric in each category, the ranks in table 4.32 are obtained.

TABLE 4.32 METRIC RANKS IN THE DIFFERENT CATEGORIES OF GLIOBLASTOMA GRN INFERRED USING WGCNA BASED ON TOP 100 GENES IDENTIFIED BY EACH METRIC

	200 or fewer category rank	**201-400 category rank**	**401-600 category rank**	**601-800 category rank**	**More than 800 category rank**	**Ranking totals**	**Overall rank**
Weighted Betweenness	1st	1st	1st	1st	1st	5	1st
Degree Centrality	1st	1st	2nd	2st	2nd	8	2nd
Weighted Degree	1st	1st	2nd	2nd	2nd	8	2nd
Weighted Closeness	1st	1st	2nd	2nd	4th	10	4th
Eigenvector Centrality	1st	1st	2nd	2nd	4th	10	4th
Combined Metric	1st	1st	2nd	2nd	4th	10	4th

Weighted betweenness is again the best performing metric, followed by degree and weighted degree. A similar situation to that of using just the top 20 genes identified by each metric occurs; some of the metrics do not identify any genes, thus having the same ranks for some of the glioblastoma categories. Finally, taking just the scores for the top 500 genes identified by each metric in each category, the ranks in table 4.33 are obtained.

TABLE 4.33 METRIC RANKS IN THE DIFFERENT CATEGORIES OF GLIOBLASTOMA GRN INFERRED USING WGCNA BASED ON TOP 500 GENES IDENTIFIED BY EACH METRIC

	200 or fewer category rank	**201-400 category rank**	**401-600 category rank**	**601-800 category rank**	**More than 800 category rank**	**Ranking totals**	**Overall rank**
Weighted Betweenness	1st	1st	1st	1st	1st	5	1st
Weighted Closeness	3rd	3rd	4th	2nd	2nd	14	2nd
Eigenvector Centrality	2nd	2nd	4th	5th	2nd	15	3rd
Weighted Degree	4th	4th	2nd	4th	4th	18	4th
Degree Centrality	4th	4th	3rd	5th	4th	20	5th
Combined Metric	6th	4th	6th	3rd	6th	25	6th

Whilst the weighted betweenness centrality is the best performing metric again, there are clear differences between the performance of the metrics. Weighted closeness is the second best performing metric, just out-performing the eigenvector centrality. The combined metric is the worst performing metric overall.

From these results, the weighted betweenness is the metric best suited to identifying glioblastoma genes of interest in the GRNs inferred for the different categories of glioblastoma.

4.9 Discussion

There are a number of areas in this chapter that merit discussion. The first of these is the combined metric gene rankings in each category. There are 57 unique entries in the five combined metric top 20 ranked gene lists. This represents quite a high proportion of genes being unique to one evolutionary category, and also that there are more unique entries in the top 20 ranked lists than common ones. This high number of unique genes suggests that the five categories of network are very different. This is especially relevant if previous static studies are considered that grouped all the samples in a data set together and constructed one network to represent their behaviour, it is quite clear from this study how different the behaviour of sample at different evolutionary stages is.

There are 14 genes that appear in two lists. 9 of these appear in the first two categories, suggesting that these two categories are the most similar based on the top ranking genes. These two categories are the two with the lowest survival days, so a reasonable presumption would be to suggest that these 9 genes play a role in the later stages of glioblastoma. It is also worth noting the very high correlation that these two categories have for common genes, 0.75, which is the highest correlation for any two categories. This again suggests that these two categories are similar. 5 genes appear in three lists, including SNAP91. Their presence in lists across the evolutionary stages would suggest they are involved in processes that are common throughout the glioblastoma life cycle, and that they do not play such an important role in the specific processes related to the evolution of glioblastoma.

The second area of focus is the biological relevance of these results. The identification of genes previously identified in a number of publications using a solely graph theory approach shows the biological relevance of this approach. Genes such as SNAP91, EPHB6, PRKCZ, CPNE6, HPCAL4 and SH3GL2 have been identified without any prior biological knowledge or bias. A number of glioblastoma signature genes were identified across the evolutionary categories using the metrics; however there was no correlation between the identification of these as being high ranking and the evolutionary category. This corresponds to the study by Verhaak et al that showed the four glioblastoma subtypes had a narrow survival range, as from this study, it cannot be concluded that one glioblastoma subtype can be significantly associated with any of the categories of network. The findings of this study suggest that high network centrality scores for specific genes may be a better indicator of glioblastoma survival time, than subtype classification.

The performance of a number of graph theory metrics was compared, based on their ability to identify genes in the different categories of GRN ranked by the text-mining tool Génie. Across all of the different categories of GRN, weighted betweenness centrality was shown to be the best performing metric at identifying these ranked genes. This might be a slightly surprising result; degree centrality is the commonly applied graph theory metric for analysis, and the number of connections that a node has is often thought of as being an indicator of the importance of a node. However, the concept of nodes that have a high betweenness acting as broker nodes has been noted, and in the context of a GRN where gene regulatory processes are likely to involve a number of genes, this perhaps should not be considered such a surprise.

The results in this chapter highlight some of the problems of applying an existing network inference to infer different categories of GRN from a microarray dataset. Whilst WGCNA has been shown to perform well at inferring one GRN from a microarray dataset, there are some issues with applying it to infer different categories of network from the same dataset. Firstly, whilst 57 out of the 100 genes in the five top 20 ranked gene lists are unique, there are 43 genes that are common to two or more glioblastoma categories. This represents quite a deal of overlap, and is not in concordance with the biology; such a high overlap amongst the categories is not expected. Email PC1 in the appendix from Dr Carmel McConville, Senior Lecturer in Cancer Sciences, notes this issue when applied to a different dataset, the underlying problem is the same here. Furthermore, looking at the composition of the largest unique cliques, there are 76 genes that are common to the largest unique cliques in all five networks, this again goes against the biology. Additionally, the identification of a number of genes being common to multiple survival categories implies that they are involved in processes common to the glioblastoma life cycle, and are not involved in processes specific to one survival category. This does not provide much assistance to biologists and clinicians wishing to interrogate microarray data to identify genes that are of interest in specific disease stages.

Secondly, the gene expression levels of the genes that are identified as being high ranking in the categories do not vary a great deal between the different categories of samples that they are taken from. This can be attributed to the method that WGCNA employs to assign the correlation scores to the networks; essentially it is looking for the greatest variability in expression level within the categories of the samples, and is not taking the expression level of the gene in the other categories of the samples into account. This is something that was noted by biologists, see email PC1, who immediately noticed this lack of variability in the expression levels of the highly ranked genes, across the categories. Leading on from this, a network inference method that takes into account the different gene expression levels of the genes across the different categories is of use; in the following chapter, a method that specifically addresses this is introduced and applied to this glioblastoma dataset.

4.10 Conclusions

To conclude, the work in this chapter has addressed the first two objectives of the work laid out in the introductory chapter. Firstly, an existing network inference technique, WGCNA, has been applied to a glioblastoma microarray dataset to infer GRNs representing different survival categories based on survival time; this was done in order to investigate the feasibility of using an existing network inference method for the purpose of inferring GRNs for different categories from a single microarray dataset. One of the issues of using an existing network inference approach is highlighted, namely that gene expression levels across the categories are not taken into account. Secondly, different metrics were used to identify genes of interest in the GRNs, and the performance of these metrics was scored and then compared, thus allowing a comparison of the ability of different metrics to identify glioblastoma genes of interest.

However, two main problems were identified with the results. A number of genes were identified as being common to both the cliques, and the highly ranked genes across the different glioblastoma categories. Additionally, the gene expression levels of the top ranked genes did not show much variance across the survival categories. These two findings are an issue as they do not agree with the feedback from biologists; it would be expected that different genes would be important in the different survival categories, and the genes of interest are those whose expression levels vary significantly across the survival categories. In the next chapter, a novel Z score based method based on the discretisation inference method developed by Vass et al [90] will be introduced, with the aim of resolving these issues by implementing a network inference that takes into account gene expression levels across different categories.

Chapter 5

Construction of Glioblastoma Survival Category GRNs using Novel Inference Method

In this chapter, an alternative novel network inference method that takes into account gene expression levels in different categories in a microarray dataset is developed, and applied to the glioblastoma dataset used in the previous chapter. It was noted in the previous chapter that the correlation-based WGCNA method used only identified genes with a high variance in their values within the categories. Whilst this is suited to the purpose of inferring a single GRN across a whole dataset, it does not seem as well suited to the purpose of inferring GRNs for different categories within a microarray dataset, and the subsequent identification of genes of interest in these different categories of GRN.

An approach that is capable of distinguishing whether genes are significantly expressed in one category of disease compared to their values in the other categories, is better suited for the purposes of this work, as suggested in email PC2 from Dr Andrew Peet, Reader in Paediatric Oncology; and more explicitly, in addressing the aim of this thesis which is to aid in the investigation of progression and evolution of tumours. Identifying genes in particular categories of disease, namely different types of cancer, has the potential to allow a better understanding of how a tumour evolves and progresses.

5.1 Novel GRN Inference Method

Some of the issues with using an existing network inference method for the inference of multiple GRNs from a single microarray dataset were highlighted in the previous chapter.

With this in mind, a novel absolute Z score network inference method is now introduced and applied to the glioblastoma dataset. This specifically addresses the first objective of the work; to design and implement a transferable method for inference of multiple GRNs from a single microarray dataset.

This method introduced here builds on the discretised approach by Vass et al detailed in chapter 3. As with the discretised method, the gene expression array levels are transformed into Z scores. However, unlike the discretised approach, these Z scores are not discretised into either a 0 or 1, and instead the Z score is transformed into an absolute value. The array of these absolute values is multiplied by the transpose of this array, resulting in a matrix of *n* by *n* dimensions, where *n* is the number of genes in the microarray dataset. This matrix is then scaled, by dividing all entries by the highest entry, resulting in a weighted adjacency matrix with values between 0 and 1. A weighted graph is inferred from this matrix; only the top 0.5% edges based on edge weight are then retained. Keeping only the top 0.5% of edges has two benefits; it results in a network with a scale-free topology, and keeping only the top 0.5% of edges based on edge weight minimises the effect of noise in the microarray data making it more probable that biologically meaningful interactions are maintained.

This network inference approach has two specific advantages over existing network inference techniques that are directly applicable to this work. Firstly, there is no minimum number of samples required in order to infer a network. This is an important consideration, as a number of microarray datasets only have a limited number of samples per category. Using the example of the WGCNA network inference technique applied in the last chapter, this method requires a minimum of 12 samples in order to infer a network. The proprietary retinoblastoma dataset that is introduced in chapter 6 has two categories where the number of samples are less than 12, highlighting this constraint. Secondly; the discretised approach developed by

Vass is also highly influenced by the number of samples. The discretised 0 or 1 nature of the scores means that the value of network metrics such as weighted degree centrality can only be one of a finite selection of values. In practice, this results in a great deal of genes having the same scores for particular metrics, and in fact means that these metrics cannot be used to discriminate between genes.

In this chapter, one GRN will be constructed for each evolutionary category detailed in the previous chapter, using the absolute Z score based network inference technique. This mirrors the WGCNA method in the networks that are constructed, and allows direct comparisons between an established method, WGCNA, and the novel method introduced and applied here.

5.2 Network Level Metrics

As was the case in the last chapter, before looking at the individual genes in the networks, the network level metrics will be calculated and analysed. The same metrics will be used again; weighted degree assortativity, degree asssortativity, diameter, and network clustering. The largest unique cliques will again be calculated for each of the networks, and the genes that comprise these cliques will be investigated for both enrichment and the presence of glioblastoma genes of interest using Génie and the list of glioblastoma subtype genes. The network level metric scores for the GRN categories are shown in table 5.1 below.

TABLE 5.1 NETWORK LEVEL SCORES GLIOBLASTOMA GRNS INFERRED USING THE NOVEL INFERENCE METHOD

Metric	**200 or less category**	**201 - 400 category**	**401 - 600 category**	**601 - 800 category**	**801+ category**
Weighted Degree Assortativity	-0.123	-0.147	-0.360	-0.323	-0.364

Degree Assortativity	-0.130	-0.134	-0.358	-0.319	-0.363
Diameter	2.030	1.889	1.143	1.638	0.934
Network Clustering	0.359	0.400	0.326	0.468	0.237
Largest Clique Size	94	144	118	217	122

There again appears to be no clear pattern between any of the network level scores and the evolutionary category. This is again suggestive that survival time cannot be associated with any network level metric behaviour. The value of the diameter should be noted here for two reasons; in the correlation based approach it was greatest in the highest survival category whilst here the exact opposite is the case, secondly it is also highest here in the lowest survival category. The diameter can be an indicator of how easily genes are able to communicate in a network; the low relative value here might be indicate genes in the highest survival category are able to communicate with each other much easier than in the lowest survival category. The network clustering is also smallest in the highest evolutionary category, as are the weighted degree and degree assortativity values. The most important observations to be made about the networks on a network level could possibly be made about the size and composition of the largest unique cliques in the networks. This is examined in depth in the following section.

5.3 Largest Clique Calculation and Analysis

As noted above, the size and composition of the largest unique cliques in the different categories can be informative. A large number of common genes in the cliques across all the networks would suggest that the same genes are involved in important cellular processes across the glioblastoma life cycle. As well as identifying common and unique genes in these cliques, it is informative to see whether these genes have been previously identified as being either glioblastoma subtype signature genes or appear on the list of Génie glioblastoma genes. Finally, the list of genes that comprise the largest unique clique in each network can be checked for enrichment using the GenesetDB website. Table 5.2 below shows whether genes in each cliques have been identified as glioblastoma subtype signature genes, or in the ranked Génie glioblastoma gene list. On the following page, the Venn diagram shows the number of unique genes in each category, and how many genes are common to two or more categories.

TABLE 5.2 GLIOBLASTOMA GENES OF INTEREST IN LARGEST UNIQUE CLIQUES IN THE GLIOBLASTOMA GRNS INFERRED USING THE NOVEL INFERENCE METHOD

Category	**Size of clique**	**Proneural Subtype Genes**	**Mesenchymal Subtype Genes**	**Neural Subtype Genes**	**Classical Subtype Genes**	**Number of Génie genes and score**
200 or less	94	1	0	12	1	5, score of 582
201 – 400	144	2	8	0	8	1, score of 122
401- 600	118	0	0	13	1	3, score of 418
601-800	217	1	17	3	7	1, score of 51
More than 800	122	4	1	3	6	5, score of 920

The first observation is the presence of fewer proneural subtype genes in these cliques compared to those in the last chapter. In contrast to those cliques in the last chapter, there are Mesenchymal subtype genes in these cliques, and also a far greater abundance of classical subtype genes as well. This clearly suggests that these two methods vary in terms of the cliques they identify. Another observation to make is the subtype ‘profile’ of these cliques, the cliques in the last chapter all had a strong neural presence; this is not the case here. Whilst the cliques in the 0-200 and 401-600 categories have a strong neural profile, this is not the case with the other cliques. The 201-400 category clique has a joint Mesenchymal/classical profile, the 601-800 clique a strong Mesenchymal/medium classical subtype profile, and the 801 and more clique a weak mixed subtype profile.

Marginally fewer glioblastoma genes of interest are identified overall in the cliques above, 104 in total in table 5.2, compared to those in the last chapter, 106 in total in table 4.4. However, more genes in the cliques in table 5.2 are present in the glioblastoma list generated by Génie; 15 genes compared to only 3 genes in the cliques in table 4.4. As noted before, the Génie list is of more interest as it allows genes to be scored based on their relevance to glioblastoma. It should also be remembered that there is a high degree of repetition in the genes that comprise the cliques shown in table 4.4.

Starting with the clique in the GRN for the 200 or fewer survival days glioblastoma category, 19 out of the 94 genes, 20.21%, have been previously identified as being of interest for glioblastoma; 14 subtype genes, and 5 genes identified by Génie. The top ten enrichment results for the genes that comprise this clique are shown in table A.12 of the appendix. The 6th result should be noted; SMAD7 has been suggested as being involved in glioblastoma cell proliferation, and is noted in a recent study by Eichhorn et al [118] as being involved in TGF-B signalling in glioblastoma.

19 out of the 144 genes, 20.21%, in the clique in the GRN for the 201-400 survival days glioblastoma category have been previously identified as being of interest for glioblastoma; 18 subtype genes, and 1 gene identified by Génie. Table A.13 of the appendix shows the top 10

results for the enrichment of the genes that comprise this clique. The 2nd result detailing ErbB1 is of note; the pathway of this gene was the 8th enrichment result for the list of top ranked genes in the 201-400 network inferred using WGCNA, and is the subject of the previously mentioned study by Gilbertson et al [116].

17 out of the 118 genes, 14.41%, in the clique in the GRN for the 401-600 survival days glioblastoma category have been previously identified as being of interest for glioblastoma; 14 subtype genes, and 3 genes identified by Génie. Table A.14 of the appendix shows the top 10 results for the enrichment of the genes that comprise this clique. Unlike the previous two categories, none of the enrichment results stand out.

29 out of the 217 genes, 13.36%, in the in the GRN for the 601-800 survival days glioblastoma category have been previously identified as being of interest for glioblastoma; 28 subtype genes, and 1 gene identified by Génie. Table A.15 of the appendix shows the top 10 results for the enrichment of the genes that comprise this clique. As was the case with the previous category, none of the enrichment results stand out.

19 out of the 122 genes, 15.57%, in the 801 or more clique have been previously identified as being of interest for glioblastoma; 14 subtype genes, and 5 genes identified by Génie. Only three results are returned by GeneSetDB for this combination of genes, suggest that this combination of genes is either not associated with any particular biological feature, or is a novel finding. The three results returned are shown in table A.16 of the appendix.

In contrast to the distribution of genes in the largest unique cliques in the last chapter, there are no genes common to all five cliques. The 200 or less, 401-600, and 801 categories are comprised of unique genes. Of the total 695 genes in all of the cliques, 650 are unique, with only 45 shared genes that belong to the 201-400 and 601-800 categories. This high proportion of unique genes and very few shared genes suggests that the cellular processes are very different in the cliques in the different categories. It might be expected that different cellular processes and genes are associated with different survival times; this result may be of more use in understanding the evolution and progression of tumours than the results in the last chapter that implied there were a high number of cellular processes common to the different categories. The Venn diagram in figure B.3 in the appendix shows the distribution of unique and shared genes in the largest cliques

Having focused on the network level metrics and cliques, in the following sections the focus is on the node-level metrics in each category.

5.4 Node Level Metrics

As before, a number of metrics for each node are calculated; weighted degree, degree, weighted betweenness, weighted closeness, eigenvector centrality, and the combined metric ranking based on weighted degree, weighted closeness, and weighted betweenness. For each category of GRN, a table of the top 20 ranked genes for the combined metric is presented. Gene set enrichment and the presence of subtype signature and glioblastoma candidate genes will be investigated for the combined list, and also whether these genes have been identified in previous glioblastoma studies.

The list of ranked glioblastoma genes from the text-mining tool Génie is used again to assign scores to the genes identified by the different metrics. The scores from the glioblastoma gene list generated by Génie of the top 20, 100, and 500 ranking genes for each metric will be used to compare the performance of the metrics in each category of glioblastoma GRN.

5.4.1 200 or Fewer Survival Days Category

Table 5.3 below shows the top 20 ranked genes for the combined metric in the GRN for the 0-200 survival days glioblastoma category. None of these genes in the table appear in the list of 299 glioblastoma genes generated by Génie. Four of the genes in the list are glioblastoma subtype signature genes. The genes P4HA2 and LOXL1 are neural subtype signature genes, the gene FHOD3 is a proneural subtype signature gene, and the gene SH3BP5 is a Mesenchymal subtype signature gene. None of the genes that appear in the top 20 ranked list appear in any of the top 20 ranked for the other categories. As was the case with the largest unique cliques, this suggests that there are clearly distinguishable differences between the survival categories in terms of the genes that are playing important roles in cellular processes.

Only one result is returned for this set of genes by GeneSetDB, shown in table A.17 of the appendix. This implies that either this combination of genes has not been identified as being involved in significant biological processes, or that this is a novel finding.

TABLE 5.3 TOP 20 RANKED GENES FOR COMBINED METRIC IN 0 – 200 SURVIVAL DAYS CATEGORY GRN INFERRED USING NOVEL INFERENCE METHOD

Gene	***Combined Metric Rank***	***Gene***	***Combined Metric Rank***
ERLIN1	1	LOXL1	11
IGF2R	2	FZD2	12
HMX1	3	SH3BP5	13
SLC22A17	4	TCF3	14
P4HA2	5	ZNRF4	15
SPAG4	6	FHOD3	16

TXNDC9	7	DNASE1L1	17
TIMM23	8	STAR	18
AMBP	9	RCC1	19
ANXA2P2	10	ZNF187	20

Table 5.4 below shows the Spearman Rank correlation scores for the rankings assigned to the genes by each metric in the 0 – 200 survival days GRN inferred using the novel method.

TABLE 5.4 SPEARMAN RANK CORRELATION SCORES OF RANKINGS ASSIGNED BY THE METRICS TO THE GENES IN 0 – 200 SURVIVAL DAYS CATEGORY GRN INFERRED USING NOVEL INFERENCE METHOD

	Degree Centrality	**Weighted Degree**	**Eigenvector Centrality**	**Closeness Centrality**	**Betweenness Centrality**
Degree Centrality	1	0.999	0.853	0.869	0.857
Weighted Degree	0.999	1	0.850	0.862	0.853
Eigenvector Centrality	0.853	0.850	1	0.853	0.694
Closeness Centrality	0.869	0.862	0.853	1	0.831
Betweenness Centrality	0.857	0.853	0.694	0.831	1

From the table above, it can be seen that all of metrics have a positive correlation greater than 0.69; if the correlation between eigenvector centrality and betweenness centrality is discounted then all the other correlations have positive correlation strength of at least 0.83. These very strong correlations imply that there is a substantial overlap in the genes that the metrics rank highly, as was the case for a lot of the metrics in the last chapter. Therefore, we would expect that there will be quite a lot of overlap in the genes that these metrics identify.

In the last chapter, a scoring system was used to compare the performance of the different metrics in identifying glioblastoma genes of interest from a ranked list. The same scoring system will be used in this chapter to compare the performance of the metrics. Using the same scoring system will allow comparisons between the performance of the metrics in identifying glioblastoma genes of interest in the networks inferred using the WGCNA network inference technique, and the performance of the metrics in this chapter using the novel Z score method. Starting with the top 20 ranked genes by each metric, table 5.5 below shows the scores that are obtained.

TABLE 5.5 GÉNIE SCORES AND METRIC RANKING FOR THE TOP 20 RANKED GENES FOR EACH METRIC IN 0 – 200 SURVIVAL DAYS CATEGORY GRN INFERRED USING NOVEL METHOD

Metric	**Number of Génie glioblastoma genes identified**	**Génie Score**	**Metric Rank**
Degree Centrality	1	6	1st
Weighted Degree	1	6	1st
Eigenvector Centrality	1	6	1st
Weighted Betweenness	0	0	4th
Weighted Closeness	0	0	4th
Combined Metric	0	0	4th

Weighted degree, degree, and eigenvector centrality each identify the same glioblastoma genes of interest, CTGF, on the list generated by Génie. The other metrics do not identify any if the genes on this list. Previously, the network metrics were scored using the top 20, top 500, and the whole network. This approach will be modified as well; the top 20, top 100, and top 500 genes ranked by each metric will be used to score the metrics. Applying the scoring system to the top 100 ranked genes for each metric, the results in table 5.6 below are given.

TABLE 5.6 GÉNIE SCORES AND METRIC RANKING FOR THE TOP 100 RANKED GENES FOR EACH METRIC IN 0 – 200 SURVIVAL DAYS CATEGORY GRN INFERRED USING NOVEL METHOD

Metric	**Number of Génie glioblastoma genes identified**	**Génie Score**	**Metric Rank**
Weighted Betweenness	2	426	1st
Eigenvector Centrality	3	421	2nd
Degree Centrality	2	240	3rd
Weighted Degree	2	240	3rd
Weighted Closeness	2	240	3rd
Combined Metric	2	240	3rd

Applying the scoring system to the top 100 genes identified by the metrics, weighted betweenness centrality is the best performing metric, followed by eigenvector centrality. Although eigenvector centrality identifies more genes, weighted betweenness centrality identifies genes with a higher score. Weighted betweenness centrality was identified as being the best performing metric overall in the last chapter for the correlation based network inference method. Finally, extending the scoring system to the top 500 genes, the following results shown in table 5.7 below are given.

TABLE 5.7 GÉNIE SCORES AND METRIC RANKING FOR THE TOP 500 RANKED GENES FOR EACH METRIC IN 0 – 200 SURVIVAL DAYS CATEGORY GRN INFERRED USING NOVEL METHOD

Metric	**Number of Génie glioblastoma genes identified**	**Génie Score**	**Metric Rank**
Weighted Degree	24	4170	1st
Degree Centrality	23	4133	2nd
Eigenvector Centrality	22	3366	3rd
Weighted Betweenness	19	2850	4th
Combined Metric	17	2391	5th
Weighted Closeness	14	1565	6th

In terms of the individual metrics, there are discernable differences in the performance of the metrics depending on the number of genes used. The performance of the degree centrality metric improves from 5th to 1st as the number of genes increases from 100 to 500, and the weighted betweenness centrality performance decreases from 1st to 4th. Interestingly, the performance of all the other metrics is the same for 100 genes, as it is for 500 genes.

Across the whole network, 149 out of the 299 glioblastoma genes generated by Génie are present. Whilst this represents just under half of these genes being present, 49.83%, only 24 are identified by the best performing metric, weighted degree centrality. In addition, in the top 500 ranked genes by weighted degree, only 7 out of the top 50 Génie ranked glioblastoma genes are present. This shows that the genes that Génie highlights as being important for glioblastoma are not identified by the network metrics in this category. This could be due to two reasons. Either the network inference method and metrics are not accurate in re-constructing and identifying the gene regulatory processes that take place, or these genes identified by Génie do not in fact play an important role in this evolutionary category. It should be noted that there is no survival information present with the genes identified by Génie, i.e. specific genes are not associated with specific survival time.

5.4.2 201-400 Survival Days Category

Table 5.8 shows the top 20 ranked genes for the combined metric in the GRN for the 201-400 survival days glioblastoma category. One of the genes, MAPK3, in the table appears in the list of 299 glioblastoma genes generated by Génie, ranked as the 122nd most important gene for glioblastoma. Three of the genes in the table are glioblastoma subtype signature genes. MMD and DMWD are classical subtype signature genes, and CRYZL1 is a signature gene of the Mesenchymal glioblastoma subtype. Again, none of the genes that appear in the top 20 ranked list appear in any of the top 20 ranked for the other categories. As was the case with the previous category, this suggests that different genes are playing important roles in the different survival categories.

Looking at the top 10 results for this list of genes from GenesetDB, shown in table A.18 of the appendix, the 3rd and 10th results stand out. IL-4 has been highlighted in a number of glioblastoma studies, such as that by Rahaman et al [119], and therefore suggests that this list of genes might be biologically significant.

TABLE 5.8 TOP 20 RANKED GENES FOR COMBINED METRIC IN 201 – 400 SURVIVAL DAYS CATEGORY GRN INFERRED USING NOVEL INFERENCE METHOD

Gene	*Combined Metric Rank*	*Gene*	*Combined Metric Rank*
STX12	1	ELF2	11

PHIP	2	SLC5A6	12
C1orf27	3	MMD	13
BCAT2	4	GOLGB1	14
IKBKB	5	ITFG1	15
GRB2	6	RALGAPB	16
HNRNPH2	7	DMWD	17
GTPBP3	8	FAM32A	18
MAPK3	9	PAFAH1B1	19
CRYZL1	10	MICU1	20

Table 5.9 below shows the Spearman Rank correlation scores for the rankings assigned to the genes by each metric in the 201 – 400 survival days GRN inferred using the novel method.

TABLE 5.9 SPEARMAN RANK CORRELATION SCORES OF RANKINGS ASSIGNED BY THE METRICS TO THE GENES IN 201 – 400 SURVIVAL DAYS CATEGORY GRN INFERRED USING NOVEL INFERENCE METHOD

	Degree Centrality	**Weighted Degree**	**Eigenvector Centrality**	**Closeness Centrality**	**Betweenness Centrality**
Degree Centrality	1	0.998	0.844	0.842	0.838
Weighted Degree	0.998	1	0.835	0.831	0.836
Eigenvector Centrality	0.844	0.835	1	0.853	0.643
Closeness Centrality	0.842	0.831	0.853	1	0.777
Betweenness Centrality	0.838	0.836	0.643	0.777	1

The correlations between the metric are strong, if the correlation between eigenvector centrality and betweenness centrality is discounted then all the other correlations have positive correlation strength of at least 0.77. Whilst these correlations are strong, they are not as strong as those in the last category, and as such, we would not expect there to be as much overlap as there was before in the genes that the metrics identify.

Applying the scoring system to the top 20, top 100, and top 500 genes identified by each of metrics, the metrics can be ranked. The scores for the top 20 genes identified by each metric are shown in table 5.10 below.

TABLE 5.10 GÉNIE SCORES AND METRIC RANKING FOR THE TOP 20 RANKED GENES FOR EACH METRIC IN 201 – 400 SURVIVAL DAYS CATEGORY GRN INFERRED USING NOVEL METHOD

Metric	**Number of Génie glioblastoma genes identified**	**Génie Score**	**Metric Rank**
Degree Centrality	1	178	1st
Weighted Degree	1	178	1st
Weighted Betweenness	1	178	1st
Eigenvector Centrality	1	178	1st
Combined Metric	1	178	1st
Weighted Closeness	0	0	6th

All of the metrics apart from the weighted closeness identify the gene MAPK3. Applying the scoring system to the top 100 ranked genes for each metric, the results in table 5.11 below are given.

TABLE 5.11 GÉNIE SCORES AND METRIC RANKING FOR THE TOP 100 RANKED GENES FOR EACH METRIC IN 201 – 400 SURVIVAL DAYS CATEGORY GRN INFERRED USING NOVEL METHOD

Metric	**Number of Génie glioblastoma genes identified**	**Génie Score**	**Metric Rank**
Weighted Betweenness	3	568	1st
Degree Centrality	2	271	2nd
Weighted Degree	2	271	2nd
Weighted Closeness	2	271	2nd
Combined Metric	2	271	2nd
Eigenvector Centrality	1	178	6th

Applying the scoring system to the top 100 genes identified by the metrics, the weighted betweenness centrality is the best performing metric. Three genes; MAPK3, MAPK1, and EGFR, are identified. It is worth noting that EGFR is the 3rd ranked gene by Génie for glioblastoma. Finally, extending the scoring system to the top 500 genes, the following results shown in table 5.12 below are given.

TABLE 5.12 GÉNIE SCORES AND METRIC RANKING FOR THE TOP 500 RANKED GENES FOR EACH METRIC IN 201 – 400 SURVIVAL DAYS CATEGORY GRN INFERRED USING NOVEL METHOD

Metric	**Number of Génie glioblastoma genes identified**	**Génie Score**	**Metric Rank**
Weighted Betweenness	11	1805	1st
Combined Metric	10	1443	2nd
Weighted Closeness	9	1308	3rd
Degree Centrality	8	1186	4th

Weighted Degree	8	1186	4th
Eigenvector Centrality	3	358	6th

The weighted betweenness centrality out-performs the other metrics when the scoring is applied to the top 500 genes identified by each metric. 2 of the top 20 ranked genes by Génie for glioblastoma, EGFR ranked 3rd and IL13RA2 ranked 17th are amongst the top 500 genes ranked by weighted betweenness centrality. This is quite a significant result in terms of the biology, and implies that weighted betweenness centrality is a good metric to use for identifying biologically significant results. The combined metric performs well due to the strong performance of the weighted betweenness centrality. Apart from eigenvector centrality, the other metrics perform reasonably well.

Across the whole network, 115 out of the 299 glioblastoma genes generated by Génie are present, 38.46%. This represents fewer genes identified than in the last category; as the survival time increases it might be expected that fewer of the genes identified by Génie might be present, a trend that fits in with the result observed.

5.4.3 401-600 Survival Days Category

Table 5.13 below shows the top 20 ranked genes for the combined metric in the GRN for the 401-600 survival days glioblastoma category. Three of the genes in the table, FOXM1, TNFRSF10B, and CDKN1A, appear in the list of 299 glioblastoma genes generated by Génie. There is only one glioblastoma subtype signature gene in the list, KCNF1, associated with the classical glioblastoma subtype. As with the two previous categories, none of the genes that appear in the table are present in the top 20 rankings for any of the other categories.

24 results are returned by GeneSetDB for this set of genes, of which 17 relate to the drug/chemical class. If these 17 results are excluded as they are not relevant to the purposes of the work, 7 results are left which are shown in table A.19 of the appendix. Of interest are the two results identifying TP53 and p53. These two genes have been highlighted in glioblastoma studies, such as those by Rasheed et al [120], and Zheng et al [121].

TABLE 5.13 TOP 20 RANKED GENES FOR COMBINED METRIC IN 401 – 600 SURVIVAL DAYS CATEGORY GRN INFERRED USING NOVEL INFERENCE METHOD

Gene	*Combined Metric Rank*	*Gene*	*Combined Metric Rank*
FOXM1	1	POLDIP2	11
HMGB3	2	TNFRSF10B	12
BSDC1	3	CDKN1A	13
APOC2	4	DDB2	14
MRPL4	5	SAP130	15
GM2A	6	BOK	16
PKN2	7	ORC5	17
TAC3	8	SRR	18
LILRB1	9	KCNF1	19
TMEM177	10	TULP3	20

Table 5.14 below shows the Spearman Rank correlation scores for the rankings assigned to the genes by each metric in the 401 – 600 survival days GRN inferred using the novel method.

TABLE 5.14 SPEARMAN RANK CORRELATION SCORES OF RANKINGS ASSIGNED BY THE DIFFERENT METRICS TO THE GENES IN 401 – 600 SURVIVAL DAYS CATEGORY GRN INFERRED USING NOVEL INFERENCE METHOD

	Degree Centrality	**Weighted Degree**	**Eigenvector Centrality**	**Closeness Centrality**	**Betweenness Centrality**
Degree Centrality	1	0.998	0.961	0.906	0.836
Weighted Degree	0.998	1	0.956	0.898	0.832
Eigenvector Centrality	0.961	0.956	1	0.934	0.796
Closeness Centrality	0.906	0.898	0.934	1	0.833
Betweenness Centrality	0.836	0.832	0.796	0.833	1

There is a positive correlation strength of at least 0.79 between all of the metrics. This implies that there is significant overlap in the genes ranked highly by the different metrics. Applying the scoring system to the top 20, top 100, and top 500 genes identified by each of metrics, the metrics can be ranked. The scores for the top 20 genes identified by each metric are shown in table 5.15 below.

TABLE 5.15 GÉNIE SCORES AND METRIC RANKING FOR THE TOP 20 RANKED GENES FOR EACH METRIC IN 401 – 600 SURVIVAL DAYS CATEGORY GRN INFERRED USING NOVEL METHOD

Metric	**Number of Génie glioblastoma genes identified**	**Génie Score**	**Metric Rank**
Degree Centrality	3	469	1st
Weighted Closeness	3	469	1st
Eigenvector Centrality	3	469	1st

Combined Metric	3	469	1st
Weighted Degree	2	313	5th
Weighted Betweenness	1	193	6th

Four of the metrics; weighted closeness, degree, eigenvector centrality and the combined metric, both identify three genes, FOXM1, TNFRSF10B, and CDKN1A, which appear on the glioblastoma genes of interest list generated by Génie. Weighted degree identifies two of these genes, FOXM1 and CDKN1A, and weighted betweenness identifies FOXM1. Applying the scoring system to the top 100 ranked genes for each metric, the results in table 5.16 below are given.

TABLE 5.16 GÉNIE SCORES AND METRIC RANKING FOR THE TOP 100 RANKED GENES FOR EACH METRIC IN 401 – 600 SURVIVAL DAYS CATEGORY GRN INFERRED USING NOVEL METHOD

Metric	**Number of Génie glioblastoma genes identified**	**Génie Score**	**Metric Rank**
Weighted Closeness	6	1310	1st
Degree Centrality	6	1152	2nd
Weighted Degree	6	1152	2nd
Weighted Betweenness	4	764	4th
Combined Metric	4	764	4th
Eigenvector Centrality	4	611	6th

Applying the scoring system to the top 100 genes identified by the metrics, the weighted closeness centrality metric is the best performing, identifying 6 genes. Degree and weighted degree also identify six genes, but these are lower ranked. The other three metrics each identify four genes. Finally, extending the scoring system to the top 500 genes, the following results shown in table 5.17 below are given.

TABLE 5.17 GÉNIE SCORES AND METRIC RANKING FOR THE TOP 500 RANKED GENES FOR EACH METRIC IN 401 – 600 SURVIVAL DAYS CATEGORY GRN INFERRED USING NOVEL METHOD

Metric	**Number of Génie glioblastoma genes identified**	**Génie Score**	**Metric Rank**
Degree Centrality	23	3979	1st
Weighted Degree	23	3979	1st
Combined Metric	20	3457	3rd
Weighted Betweenness	18	3404	4th
Weighted Closeness	17	3162	5th
Eigenvector Centrality	18	2957	6th

The degree and weighted degree metrics are the joint best performing metrics, each identifying the same 23 glioblastoma genes. Previously the performance of the weighted betweenness centrality, the 4th best performing metric, at identifying highly ranked glioblastoma genes was noted. This time, the weighted closeness centrality, the 5th best performing gene this time, performs impressively. This metric identifies 3 of the top 15 ranked genes by Génie for glioblastoma amongst its top 500 ranked genes; MGMT ranked 2nd, PTEN ranked 5th, and BIRC5 ranked 13th. This result implies that weighted closeness centrality is a good metric to use for identifying biologically significant genes. In general the metrics perform strongly for this category; and identify a number of glioblastoma genes.

Across the whole network, 126 out of the 299 glioblastoma genes generated by Génie are present, 42.14%. This represents more genes identified than in the last category; as is contrary to the observation for the last category, we might expect fewer genes to be identified as the survival time increases, instead of more genes compared to the last category.

5.4.4 601-800 Survival Days Category

Table 5.18 below shows the top 20 ranked genes for the combined metric in the GRN for the 601-800 survival days glioblastoma category. The 7th ranked gene, PTK2B, appears in the list of 299 glioblastoma genes generated by Génie. One glioblastoma subtype gene is present in the table, EPS15, which is Mesenchymal subtype signature gene. As before, none of the genes that appear in the top 20 ranked list appear in any of the top 20 ranked for the other categories.

No results are returned for this set of genes by GeneSetDB. This implies that either this combination of genes has not been identified as being involved in significant biological processes, or that this is a novel finding.

TABLE 5.18 TOP 20 RANKED GENES FOR COMBINED METRIC IN 601 – 800 SURVIVAL DAYS CATEGORY GRN INFERRED USING NOVEL INFERENCE METHOD

Gene	***Combined Metric Rank***	***Gene***	***Combined Metric Rank***
CLSTN3	1	PEX16	11
NPY1R	2	FRY	12
CBFB	3	MDN1	13
PTK2B	4	EML4	14
HMG20A	5	ARHGEF9	15
EPS15	6	GLRB	16
ZDHHC11	7	CDK5R2	17

CAMKK2	8	FAM190B	18
HNRPDL	9	CIT	19
GFRA2	10	FGF14	20

Table 5.19 below shows the Spearman Rank correlation scores for the rankings assigned to the genes by each metric in the 601 – 800 survival days GRN inferred using the novel method.

TABLE 5.19 SPEARMAN RANK CORRELATION SCORES OF RANKINGS ASSIGNED BY THE METRICS TO THE GENES IN 601 – 800 SURVIVAL DAYS CATEGORY GRN INFERRED USING NOVEL INFERENCE METHOD

	Degree Centrality	**Weighted Degree**	**Eigenvector Centrality**	**Closeness Centrality**	**Betweenness Centrality**
Degree Centrality	1	0.997	0.872	0.844	0.746
Weighted Degree	0.997	1	0.870	0.835	0.743
Eigenvector Centrality	0.872	0.870	1	0.941	0.615
Closeness Centrality	0.844	0.835	0.941	1	0.693
Betweenness Centrality	0.746	0.743	0.615	0.693	1

The correlation scores between the metrics are weaker this time. Whilst some of the correlations are strong and as a result there will still be some overlap between the genes identified by the metrics, we would not expect there to be as much overlap as previously when the correlations were stronger.

Applying the scoring system to the top 20, top 100, and top 500 genes identified by each of metrics, the metrics can be ranked. The scores for the top 20 genes identified by each metric are shown in table 5.20 below.

TABLE 5.20 GÉNIE SCORES AND METRIC RANKING FOR THE TOP 20 RANKED GENES FOR EACH METRIC IN 601 – 800 SURVIVAL DAYS CATEGORY GRN INFERRED USING NOVEL METHOD

Metric	**Number of Génie glioblastoma genes identified**	**Génie Score**	**Metric Rank**
Weighted Betweenness	2	333	1st
Weighted Closeness	2	333	1st
Degree Centrality	1	51	3rd
Weighted Degree	1	51	3rd
Eigenvector Centrality	1	51	3rd
Combined Metric	1	51	3rd

The weighted closeness and weighted betweenness both identify the same two genes, TNC ranked 18th and the previously mentioned PTK2B ranked 249th by Génie. The other metrics all identify one gene, PTK2B. Applying the scoring system to the top 100 ranked genes for each metric, the results in table 5.21 below are given.

TABLE 5.21 GÉNIE SCORES AND METRIC RANKING FOR THE TOP 100 RANKED GENES FOR EACH METRIC IN 601 – 800 SURVIVAL DAYS CATEGORY GRN INFERRED USING NOVEL METHOD

Metric	**Number of Génie glioblastoma genes identified**	**Génie Score**	**Metric Rank**
Weighted Closeness	4	677	1st

Degree Centrality	2	333	2nd
Weighted Betweenness	2	333	2nd
Combined Metric	2	333	2nd
Weighted Degree	1	51	5th
Eigenvector Centrality	1	51	5th

The weighted closeness is the best performing metric, identifying four glioblastoma genes, TNC and PTK2B identified before, and CTNNB1 and PEG3. The weighted degree and eigenvector centrality metrics perform badly, only identifying the same gene that they identified before in their top 20 rankings. Finally, extending the scoring system to the top 500 genes, the following results shown in table 5.22 below are given.

TABLE 5.22 GÉNIE SCORES AND METRIC RANKING FOR THE TOP 500 RANKED GENES FOR EACH METRIC IN 601 – 800 SURVIVAL DAYS CATEGORY GRN INFERRED USING NOVEL METHOD

Metric	**Number of Génie glioblastoma genes identified**	**Génie Score**	**Metric Rank**
Weighted Betweenness	15	2698	1st
Weighted Closeness	13	2148	2nd
Combined Metric	10	1745	3rd
Degree Centrality	7	1361	4th
Weighted Degree	7	1361	4th
Eigenvector Centrality	6	731	6th

Applying the scoring to the top 500 genes identified by each metric, weighted betweenness is the best performing metric, followed by the weighted closeness centrality. Whilst all of the metrics perform worse than in the previous category, weighted betweenness, weighted degree and degree identify 2 of the top 20 ranked genes by Génie for glioblastoma. TNC ranked 18th, and PTGS2 ranked 20th, are amongst the respective top 500 genes ranked by these metrics. This is still quite a significant result in terms of the biology, and implies that these metrics are able to identify biologically significant genes.

Across the whole network, 102 out of the 299 glioblastoma genes generated by Génie are present, 30.43%. This is the lowest amount of genes identified any of the categories; whilst this was not the case with the previous category, as the survival time increases it might be expected that fewer of the genes identified by Génie might be present.

5.4.5 More than 800 Survival Days Category

Table 5.23 below shows the top 20 ranked genes for the combined metric in the GRN for the more than 800 survival days glioblastoma category. Two genes, NR2E1 and NBN, appear in the list of 299 glioblastoma genes generated by Génie. There is only one glioblastoma subtype gene in the table; as well being ranked as the 220th most important glioblastoma gene by Génie, the 15th ranked gene NR2E1 is also a proneural subtype signature gene. Again, none of the genes that appear in the top 20 ranked list appear in any of the top 20 ranked for the other categories.

49 results are returned GenesetDB for this set of genes, of which 34 relate to the drug/chemical class, which are not informative for our purposes. This leaves 15 results, the top 10 of which are shown in table A.20 of the appendix. The 6th result should be noted, as RAC1 has been highlighted in a study by Chan et al [122] investigating glioblastoma cell invasion.

TABLE 5.23 TOP 20 RANKED GENES FOR COMBINED METRIC IN MORE THAN 800 SURVIVAL DAYS CATEGORY GRN INFERRED USING NOVEL INFERENCE METHOD

Gene	*Combined Metric Rank*	*Gene*	*Combined Metric Rank*
ABI1	1	TIAM1	11
AKR1C3	2	RASSF4	12
AKR1C1	3	JAK1	13
OSBPL11	4	NR2E1	14
ADCYAP1R1	5	RABL3	15
CALHM2	6	ARPC1A	16
PARD3	7	NBN	17
PLAC8	8	ARHGAP12	18
ABLIM3	9	PACSIN3	19
TJP2	10	UPF1	20

Table 5.24 below shows the Spearman Rank correlation scores for the rankings assigned to the genes by each metric in the more than 800 survival days GRN inferred using the novel method.

TABLE 5.24 SPEARMAN RANK CORRELATION SCORES OF RANKINGS ASSIGNED BY THE METRICS TO THE GENES IN MORE THAN 800 SURVIVAL DAYS CATEGORY GRN INFERRED USING NOVEL INFERENCE METHOD

	Degree	**Weighted Degree**	**Eigenvector Centrality**	**Closeness Centrality**	**Betweenness Centrality**
Degree	1	0.998514131	0.969827884	0.851143211	0.815157589
Weighted Degree	0.998514131	1	0.968143263	0.843257684	0.809975183

Eigenvector Centrality	0.969827884	0.968143263	1	0.867907565	0.783339875
Closeness Centrality	0.851143211	0.843257684	0.867907565	1	0.758443262
Betweenness Centrality	0.815157589	0.809975183	0.783339875	0.758443262	1

The correlations between the metrics in this category are strong, with the weakest correlation between weighted betweenness centrality and weighted closeness centrality have a strength of 0.75. The next weakest correlation is 0.78. We would expect to see substantial overlap in the genes identified by the metrics due to these strong correlations.

Applying the scoring system to the top 20, top 100, and top 500 genes identified by each of metrics, the metrics can be ranked. The scores for the top 20 genes identified by each metric are shown in table 5.25 below.

TABLE 5.25 GÉNIE SCORES AND METRIC RANKING FOR THE TOP 20 RANKED GENES FOR EACH METRIC IN MORE THAN 800 SURVIVAL DAYS CATEGORY GRN INFERRED USING NOVEL METHOD

Metric	**Number of Génie glioblastoma genes identified**	**Génie Score**	**Metric Rank**
Weighted Degree	1	267	1st
Eigenvector Centrality	1	267	1st
Combined Metric	2	262	3rd
Weighted Closeness	1	182	4th
Weighted Betweenness	1	80	5th
Degree Centrality	0	0	6th

The weighted degree and eigenvector centrality are the joint best performing metrics, despite identifying one gene less than the combined metric, which identifies the genes NR2E1 and NBN. This is due to the weighted degree and eigenvector centrality identifying EPHA2, the 33rd ranked gene by Génie. Applying the scoring system to the top 100 ranked genes for each metric, the results in table 5.26 below are given.

TABLE 5.26 GÉNIE SCORES AND METRIC RANKING FOR THE TOP 100 RANKED GENES FOR EACH METRIC IN MORE THAN 800 SURVIVAL DAYS CATEGORY GRN INFERRED USING NOVEL METHOD

Metric	**Number of Génie glioblastoma genes identified**	**Génie Score**	**Metric Rank**
Combined Metric	9	1104	1st
Weighted Betweenness	6	964	2nd
Degree Centrality	6	879	3rd
Weighted Degree	4	813	4th
Weighted Closeness	8	808	5th
Eigenvector Centrality	4	800	6th

The combined metric again identifies the most genes, and is the best performing metric this time. The weighted betweenness is the next best performing, followed by the degree centrality. Finally, extending the scoring system to the top 500 genes, the following results shown in table 5.27 below are given.

TABLE 5.27 GÉNIE SCORES AND METRIC RANKING FOR THE TOP 500 RANKED GENES FOR EACH METRIC IN MORE THAN 800 SURVIVAL DAYS CATEGORY GRN INFERRED USING NOVEL METHOD

Metric	**Number of Génie glioblastoma genes identified**	**Génie Score**	**Metric Rank**
Degree Centrality	22	2517	1st
Weighted Degree	22	2517	1st
Eigenvector Centrality	19	2326	3rd
Weighted Betweenness	20	2195	4th
Combined Metric	17	1899	5th
Weighted Closeness	18	1733	6th

Compared to the last category, the metrics identify quite a few glioblastoma genes, with the degree and weighted degree the best performing metrics. However, only 1 of the top 30 glioblastoma genes ranked by Génie are identified by any of the metrics; the gene AKT1, ranked 16th. It might be expected that as this is the longest survival category time, genes that have been identified as being highly ranked for glioblastoma might not be as prominent in the network, something that this result would suggest.

Across the whole network, 139 out of the 299 glioblastoma genes generated by Génie are present, 42.81%. This is the 2nd highest amount of genes identified in any of the categories, only fewer than the 200 or less category. It might be expected that the lowest amount of Génie genes would be present in this category, however it has been noted that highly ranked Génie genes do not figure prominently in terms of their network metrics. This suggests that those genes most associated with glioblastoma cellular processes are not as active in the network in this category, something that would be expected considering the survival time of this category.

5.5 Metric Performance

Previously, each of the metrics was ranked based on the top 20, top 100 and top 500 genes they identified for each glioblastoma GRN category. This ranking is based on the score of the identified genes in the Génie glioblastoma gene list.

The results of these metric rankings can now be presented together, allowing an analysis of which metric is the best performing overall. Assigning an equal weighting to the scores of the top 20, top 100, and top 500 ranked genes for each metric in each category, the following ranks shown in table 5.28 below are given.

TABLE 5.28 OVERALL METRIC RANKS IN THE DIFFERENT CATEGORIES OF GLIOBLASTOMA GRN INFERRED USING NOVEL INFERENCE METHOD

	200 or fewer category rank	201-400 category rank	401-600 category rank	601-800 category rank	More than 800 category rank	Ranking totals	Overall rank
Degree Centrality	2nd	3rd	1st	4th	3rd	13	1st
Weighted Degree	1st	3rd	3rd	5th	1st	13	1st
Combined Metric	5th	2nd	3rd	3rd	2nd	15	3rd
Weighted Betweenness	4th	1st	6th	1st	5th	17	4th
Weighted Closeness	6th	5th	2nd	1st	6th	20	5th
Eigenvector Centrality	2nd	6th	5th	6th	3rd	22	6th

Overall, degree and weighted degree are the best performing metrics for this dataset using the Z score network inference approach. The degree is the most widely studied and applied network metric for analysis, and it appears that for this dataset, it is the most informative overall. If the rankings for just the top 20 genes in each category are taken into consideration, the ranks shown in table 5.29 are obtained.

TABLE 5.29 METRIC RANKS IN THE DIFFERENT CATEGORIES OF GLIOBLASTOMA GRN INFERRED USING NOVEL INFERENCE METHOD BASED ON TOP 20 GENES IDENTIFIED BY EACH METRIC

	200 or fewer category rank	**201-400 category rank**	**401-600 category rank**	**601-800 category rank**	**More than 800 category rank**	**Ranking totals**	**Overall rank**
Eigenvector Centrality	1st	1st	1st	3rd	1st	7	1st
Weighted Degree	1st	1st	5th	3rd	1st	11	2nd
Combined Metric	4th	1st	1st	3rd	3rd	12	3rd
Degree Centrality	1st	1st	1st	3rd	6th	12	3rd
Weighted Closeness	4th	6th	1st	1st	4th	16	5th
Weighted Betweenness	4th	1st	6th	1st	5th	17	6th

The degree and weighted degree metrics perform strongly again, but this time their performance is bettered by the eigenvector centrality. The popular Google search engine uses a variant of the eigenvector centrality, Pagerank, and for this dataset it appears that a metric that takes into account the properties of the links of a node as well as the number of links, performs better than metrics that simply take into account the number or strength of links. The combined metric also performs well, matching the performance of the degree centrality. Taking the ranking for just the top 100 genes into account, the following ranks shown in table 5.30 obtained.

TABLE 5.30 METRIC RANKS IN THE DIFFERENT CATEGORIES OF GLIOBLASTOMA GRN INFERRED USING NOVEL INFERENCE METHOD BASED ON TOP 100 GENES IDENTIFIED BY EACH METRIC

	200 or fewer category rank	**201-400 category rank**	**401-600 category rank**	**601-800 category rank**	**More than 800 category rank**	**Ranking totals**	**Overall rank**
Weighted Betweenness	1st	1st	4th	2nd	2nd	10	1st
Weighted Closeness	3rd	2nd	1st	1st	5th	12	2nd
Combined Metric	3rd	2nd	4th	2nd	1st	12	2nd
Degree Centrality	3rd	2nd	2nd	2nd	3rd	12	2nd
Weighted Degree	3rd	2nd	2nd	5th	4th	16	5th
Eigenvector Centrality	2nd	6th	6th	5th	6th	25	6th

Weighted betweenness is the best performing metric again, jointly followed by weighted closeness and degree centrality. The eigenvector centrality is the worst performing metric this time, perhaps somewhat surprising to see the performance decrease by so much as the number of genes is increased. Finally, solely taking the network rankings for the top 500 genes into account, the following rankings shown in table 5.31 are obtained.

TABLE 5.31 METRIC RANKS IN THE DIFFERENT CATEGORIES OF GLIOBLASTOMA GRN INFERRED USING NOVEL INFERENCE METHOD BASED ON TOP 500 GENES IDENTIFIED BY EACH METRIC

	200 or fewer category rank	**201-400 category rank**	**401-600 category rank**	**601-800 category rank**	**More than 800 category rank**	**Ranking totals**	**Overall rank**
Weighted Degree	1st	4th	1st	4th	1st	11	1st
Degree Centrality	2nd	4th	1st	4th	1st	12	2nd
Weighted Betweenness	4th	1st	4th	1st	4th	14	3rd
Combined Metric	5th	2nd	3rd	3rd	5th	18	4th
Weighted Closeness	6th	3rd	5th	2nd	6th	22	5th
Eigenvector Centrality	3rd	6th	6th	6th	3rd	24	6th

Weighted degree is the best performing metric, followed by the degree centrality. The weighted betweenness still performs relatively well, but not as well as for the top 100 genes. The eigenvector centrality is again the worst performing metric.

5.6 Distribution of Genes in the Combined Metric Rankings List

As noted previously, none of the genes that appear in the top 20 combined metric rankings list for any category appear in any other. This can be seen on the Venn diagram in figure B.4 of the appendix below, showing the distribution of genes in the top 20 combined metric rankings list in the categories. Furthermore, extending this to look at the top 100 combined metric rankings lists, shown in figure B.5 of the appendix, this is also the case. This suggests that this approach is suited to identifying specific genes associated with the different glioblastoma survival times.

5.7 Comparison of Glioblastoma Results from different Network Inference Methods

Having inferred networks for the same survival categories in the glioblastoma dataset using both the WGCNA network inference method in the previous chapter, and the Z score based network inference method in this chapter, the results can be compared. More specifically, the number of ranked glioblastoma genes identified in each of the five categories of network can be compared for both network inference methods. In this section, the number and scores of the ranked glioblastoma genes identified by both the combined metric, and also the best performing metric, for the networks inferred using both of the network inference methods for each survival category are compared. Table 5.32 below shows the number and scores of ranked Génie glioblastoma genes identified in the networks inferred for the 200 or less survival days glioblastoma category.

TABLE 5.32 RANKED GLIOBLASTOMA GENES IDENTIFIED IN 200 OR LESS SURVIVAL DAYS CATEGORY GRNS INFERRED USING WGCNA AND NOVEL INFERENCE METHOD

200 or less category	Top 20 ranked genes	Top 100 ranked genes	Top 500 ranked genes
WGCNA combined metric	0	1, score of 165	4, score of 238
Z Score combined metric	0	2, score of 240	17, score of 2391
WGCNA best performing metric	0	1, score of 165	9, score of 1107
Z Score best performing metric	1, score of 6	2, score of 426	24, score of 4170

For the 200 or less survival days category, the combined metric in the GRN inferred using the Z score approach identifies more ranked glioblastoma genes 2 out of the 3 times. The other time, the combined metric fails to identify any ranked glioblastoma genes in either the network inferred using WGCNA, or the Z score approach. The best performing metric in the network inferred using the Z score approach outperforms the best performing metric in the network inferred using WGCNA all three times. Table 5.33 below shows the number and scores of ranked Génie glioblastoma genes identified in the networks inferred for the 201-400 survival days glioblastoma category.

TABLE 5.33 RANKED GLIOBLASTOMA GENES IDENTIFIED IN 201 - 400 SURVIVAL DAYS CATEGORY GRNS INFERRED USING WGCNA AND NOVEL INFERENCE METHOD

201 - 400 category	Top 20 ranked genes	Top 100 ranked genes	Top 500 ranked genes
WGCNA combined metric	0	0	5, score of 370
Z Score combined metric	1, score of 178	2, score of 271	10, score of 1443
WGCNA best performing metric	0	0	6, score of 944
Z Score best performing metric	1, score of 178	3, score of 568	11, score of 1805

This time, the combined metric in the network inferred using the Z score approach identifies more ranked glioblastoma genes all 3 times. This is also the case for the best performing metric in the network inferred using the Z score approach which outperforms the best performing metric in the network inferred using WGCNA all three times. Table 5.34 below shows the number and scores of ranked Génie glioblastoma genes identified in the network inferred for the 401-600 survival days glioblastoma category.

TABLE 5.34 RANKED GLIOBLASTOMA GENES IDENTIFIED IN 401 - 600 SURVIVAL DAYS CATEGORY GRNS INFERRED USING WGCNA AND NOVEL INFERENCE METHOD

401 - 600 category	**Top 20 ranked genes**	**Top 100 ranked genes**	**Top 500 ranked genes**
WGCNA combined metric	0	0	3, score of 215
Z Score combined metric	3, score of 469	4, score of 764	20, score of 3457
WGCNA best performing metric	1, score of 273	3, score of 610	10, score of 1895
Z Score best performing metric	3, score of 469	6, score of 1310	23, score of 3979

Again, the combined metric and best performing metrics in the network inferred using the Z score approach identify more ranked glioblastoma genes all 3 times. Table 5.35 below shows the number and scores of ranked Génie glioblastoma genes identified in the GRNs inferred for the 601-800 survival days glioblastoma category.

TABLE 5.35 RANKED GLIOBLASTOMA GENES IDENTIFIED IN 601 - 800 SURVIVAL DAYS CATEGORY GRNS INFERRED USING WGCNA AND NOVEL INFERENCE METHOD

601 - 800 category	**Top 20 ranked genes**	**Top 100 ranked genes**	**Top 500 ranked genes**
WGCNA combined metric	0	0	3, score of 563
Z Score combined metric	1, score of 51	2, score of 333	10, score of 1745
WGCNA best performing metric	2, score of 268	3, score of 488	6, score of 1036
Z Score best performing metric	2, score of 333	4, score of 677	15, score of 2698

As before, the combined metric and best performing metrics in the network inferred using the Z score approach identify more ranked glioblastoma genes all 3 times. Finally, table 5.36 below shows the number and scores of ranked Génie glioblastoma genes identified in the GRNs inferred for the more than 800 survival days glioblastoma category.

TABLE 5.36 RANKED GLIOBLASTOMA GENES IDENTIFIED IN MORE THAN 800 SURVIVAL DAYS CATEGORY GRNS INFERRED USING WGCNA AND NOVEL INFERENCE METHOD

More than 800 category	**Top 20 ranked genes**	**Top 100 ranked genes**	**Top 500 ranked genes**
WGCNA combined metric	0	0	5, score of 450
Z Score combined metric	2, score of 262	9, score of 1104	17, score of 1899
WGCNA best performing metric	0	1, score of 68	10, score of 804
Z Score best performing metric	1, score of 267	9, score of 1104	22, score of 2517

The combined metric and best performing metrics in the network inferred using the Z score approach identify more ranked glioblastoma genes all 3 times for this survival category. Overall, the combined metric in the networks inferred using the Z score approach outperform the combined metric in the networks inferred using the WGCNA method 14 out of 15 times. The other time, the combined metric fails to identify any ranked glioblastoma genes. The best performing metrics in the networks inferred using the Z score approach outperform the best performing metrics in the networks inferred using the WGCNA method all 15 times. This suggests that as the same metrics were used to identify genes in all of the networks, based on the presence of ranked glioblastoma genes, the Z score network inference method outperforms the WGCNA network inference method at identifying glioblastoma genes of interest across different survival categories in the glioblastoma dataset. This result could also suggest that highly ranked genes identified by the Z score approach that have not been previously identified as being of interest for glioblastoma might be worth investigating as potential new glioblastoma genes of interest.

5.8 Discussion

Having inferred networks for the different survival categories using the Z score absolute value approach, and subsequently analysed the results, a number of concluding observations can be made. Firstly, unique genes were identified as being prominent in the networks for each of the survival categories, using the combined metric. This is a result that is in concordance with biological principles; it would be expected that different genes would be involved in cellular processes associated with different survival times. It is also a result that biologically could be informative, as it specifically highlights different genes associated with specific survival time. This was not the case with the correlation based approach, where there was a substantial overlap amongst the 100 genes in total identified from the 5 categories, 43 in fact, suggesting that the previous approach was not the best suited to identifying specific genes associated with specific survival times. Extending the number of genes to the top 100 in each category based on the combined metric, there is still no overlap of genes, with only completely unique genes in each category.

Secondly, those genes identified as being of interest show significant variability in their gene expression levels across the different survival categories. As was the case with the identification of unique genes being prominent in the different survival categories, this is a result that observes the biological principles, differences in gene expression levels would be expected to be responsible for different gene behaviour in the different survival categories.

Furthermore, the biologists believe these results to be a much better representation of the gene regulatory processes in the different survival categories than was the case with the results presented in the last chapter.

The identification of genes previously identified in a number of publications, using a solely graph theory approach shows the biological relevance of a graph theoretical approach to microarray data. The enrichment results of these genes that relate to glioblastoma cellular processes also show the biological significance of this approach, with a number of biological processes associated with glioblastoma present in the enrichment results. Genes such as MAPK3, FOXM1, PTK2B, and NR2E1, have been identified without any prior biological knowledge or bias. As was the case with the WGCNA network inference approach, a number of glioblastoma subtype signature genes were identified across the evolutionary categories using the metrics; however there was no correlation between the identification of these as being high ranking and the evolutionary category. This is in concordance with the study by Verhaak et al that showed the glioblastoma subtypes had a narrow survival range. The results in this chapter show that glioblastoma subtype cannot be significantly associated with any of the survival categories of network, thereby suggesting that glioblastoma subtype cannot be significantly associated with survival time for this dataset. Again, the findings of the Z score network inference approach suggest that high network centrality scores for specific genes may be a better indicator of glioblastoma survival time, than subtype classification.

The distribution of the genes in the largest unique cliques is also very different to that in the last chapter. In the last chapter, there were 79 genes that were common to all five of the cliques. Using the Z score approach, no genes are common to all five cliques; three of the largest unique cliques for the categories are comprised of unique genes and the only overlap of genes between cliques are the 45 genes that appear in both the 201-400 and 601-800 categories. As was the case with the genes identified using the combined metric, it would be expected that different genes would be involved in the different cellular processes associated with different survival times, something that is largely the case with the largest unique cliques. The enrichment results for the cliques suggest that particular processes involved in glioblastoma occur for different survival categories, a result that also highlights the differences between the networks of the different categories.

5.9 Conclusions

To conclude, in this chapter a network inference method to address the first objective was proposed. This method takes into account some of the constraints of using existing network inference methods to infer multiple GRNs from a single microarray dataset highlighted in the previous chapter, and addresses these. In the next chapter, this method will be applied to two neuroblastoma microarray datasets in an attempt to gauge whether it is transferable to other cancer microarray datasets.

The second objective was also addressed; various metrics were used to identify different genes in the different categories of glioblastoma GRNs, and the performance of the metrics was compared and analysed. Unique genes were identified based on their metric scores for each glioblastoma GRN category. Additionally, these genes identified have significant variability in their expression levels across the survival categories, two results that are in keeping with observed biological phenomena. A greater number of ranked glioblastoma genes of interest were identified using the novel approach introduced in this chapter than with the WGCNA approach used in the last chapter.

Chapter 6

Analysis of Disease Stage GRNs in Two Neuroblastoma Microarray Datasets

In the previous chapter, a novel network inference method was introduced; this method was used to infer a number of GRNs for different survival categories of glioblastoma from a single microarray dataset, and graph theory metrics were applied to identify genes of interest in the different categories of GRN. In this chapter, the same novel network inference method is applied to two neuroblastoma datasets, from well-cited studies by Molenaar et al [123], and Wang et al [124]. These two datasets will be referred to as the Molenaar dataset, and the Wang dataset, respectively.

The work in this chapter addresses all three objectives specified for this work. Firstly, the novel network inference method is used to infer multiple GRNs from two different microarray datasets of the same disease; thereby investigating the transferability of the method, by applying it to two different microarray datasets of the same disease. A different type of disease microarray dataset is used than in the last chapter.

Secondly, different graph theory metrics are applied to the GRNs to identify high ranking genes. In the previous two chapters, a number of genes previously highlighted in various studies as being relevant to glioblastoma were identified through the application of graph theory metrics to the GRNs inferred. This approach is repeated here; to investigate whether this approach is also valid for other disease type microarray datasets, in this case neuroblastoma.

Thirdly, the GRN networks inferred for the disease stages that are common to both the Molenaar and Wang neuroblastoma microarray datasets are compared. This is carried out with the aim of investigating whether repeatable results are obtained across different microarrays for the same disease, and whether as a result common genes are identified with the same disease stage across both datasets.

6.1 Neuroblastoma Microarray Datasets

Having applied the Z score network inference technique to a glioblastoma dataset, in this chapter it is applied to neuroblastoma. Neuroblastoma is the most commonly diagnosed extracranial cancer in infancy; it is an embryonal tumour of the autonomic nervous system, with an incidence of 10.2 cases per million in children under 15 years of age [125]. It is highly variable; a staging system previously developed to classify neuroblastoma between 1986 and 1988, the International Neuroblastoma Staging System, INSS, has been further refined recently based on the correlation between MYCN amplification and survival [126]. Whilst the two previous chapters detailed the analysis of one dataset, in this chapter two datasets from two independent well-cited neuroblastoma studies will be used. The rationale for using two datasets is to investigate whether results can be replicated across datasets from different studies. Due to the use of greatly varying training datasets for the creation and validation of network inference techniques, replicating results using network inference techniques across different datasets is something of an issue. With the aim of investigating whether results can be replicated using the Z score network inference, the two datasets from the well-cited Molenaar et al [123], and Wang et al [124], studies will be used.

In order to apply the same inference and analysis to both datasets, the genes that are common to both of the datasets can be used. There are 4829 genes common to the Molenaar and Wang datasets. The Molenaar dataset consists of 88 samples, and the Wang dataset comprises 101 samples. Four categories of samples were used from both of these neuroblastoma datasets, following input from biologists. INSS stage was used as an initial class discriminator; stage 4 samples were then further classified based on whether the gene MYCN was amplified or not. The amplification of the gene MYCN in stage 4 neuroblastoma is an area of particular interest for many biologists researching neuroblastoma [127], following feedback from biologists this was used as part of the criteria for categorisation. Those with significant MYCN amplification were labelled stage 4M, and those without significant amplification labelled stage 4. Due to the absence of INSS stage 2 samples in the Wang dataset, all stage 2 samples in the Molenaar dataset were discarded. Likewise, all stage4S samples were also discarded from the Molenaar dataset. This left 61 samples in the Molenaar dataset, and 101 samples in the Wang dataset. Table 6.1 shows the number of samples in each category for the Verbeeg and Wang datasets.

TABLE 6.1 NUMBER OF SAMPLES IN THE COMMON NEUROBLASTOMA CATEGORIES IN THE TWO DATASETS

Category	*Number of Samples Molenaar Dataset*	*Number of Samples Wang Dataset*
Stage 1	8	28
Stage 3	13	23
Stage 4	20	30
Stage 4M	20	20

From table 6.1 above, it can be seen that the samples are spread out quite well across the categories. As with the glioblastoma dataset used previously, one analysis that highlights the value of the network inference method on the categories of data is to rank the mean expression for each gene in the categories, and see how well these rankings correlate across the four categories. This gives an idea of how homogenous the data is across the categories before the networks are constructed. This analysis also allows an initial comparison of the two datasets, to see how similar they are in terms of the expression levels of the common genes. Table 6.2 below shows the correlations for the Molenaar dataset, and table 6.3 shows the correlations for the Wang dataset. Table 6.4 shows the correlations between the common categories between both neuroblastoma datasets.

TABLE 6.2 CORRELATION SCORES OF AVERAGE GENE EXPRESSION RANKINGS MOLENAAR DATASET

Category	Stage 1	Stage 3	Stage 4	Stage 4M
Stage 1	1	0.976	0.981	0.963
Stage 3	0.976	1	0.979	0.969
Stage 4	0.981	0.979	1	0.983
Stage 4M	0.963	0.969	0.983	1

TABLE 6.3 CORRELATION SCORES OF AVERAGE GENE EXPRESSION RANKINGS WANG DATASET

Category	Stage 1	Stage 3	Stage 4	Stage 4M
Stage 1	1	0.976	0.976	0.939
Stage 3	0.976	1	0.971	0.934
Stage 4	0.976	0.971	1	0.958
Stage 4M	0.939	0.934	0.958	1

TABLE 6.4 CORRELATION SCORES OF AVERAGE GENE EXPRESSION RANKINGS BETWEEN MOLENAAR AND WANG DATASETS

Category	Wang Stage 1	Wang Stage 3	Wang Stage 4	Wang Stage 4M
Molenaar Stage 1	0.631	0.608	0.604	0.591
Molenaar Stage 3	0.616	0.588	0.588	0.590
Molenaar Stage 4	0.631	0.608	0.621	0.621
Molenaar Stage 4M	0.610	0.592	0.602	0.630

The correlations for the first two tables, table 6.2 and table 6.3, inter-dataset, are both positive and extremely strong, again showing that by simply looking at gene expression in each category it is difficult to distinguish the categories, and gain meaningful information about them. However, the correlations across the two datasets are weaker, table 6.4, showing that the datasets are somewhat different from each other. This result would suggest that the gene expression levels of the common genes are significantly different from each other, and that as a result of this, different results might be expected from the two datasets. This is in keeping with the earlier observation about repeatability of results across different datasets; differences in the datasets used leads to a lack of repeatable observations. This is something that might be also expected to be the case here.

6.2 Network Level Metrics in the Neuroblastoma GRNs

As was the case in the last chapter, before looking at the individual genes in the networks, the network level metrics will be calculated and analysed. These are shown in table 6.5 for the

Molenaar dataset, and table 6.6 for the Wang dataset. The same metrics will be used again; weighted degree assortativity, degree asssortativity, diameter, and network clustering. The largest unique cliques will again be calculated for each of the networks for the two different datasets, and the genes that comprise these cliques will be investigated for both enrichment and the presence of neuroblastoma genes of interest using Génie. The list generated by Génie for neuroblastoma is also much larger than for glioblastoma, in fact only the top 500 ranked genes for neuroblastoma will be used. The Génie list of the scoring genes for neuroblastoma is shown in table A.21 of the appendix. Unlike with glioblastoma, there are no neuroblastoma subtype gene lists to use.

TABLE 6.5 NETWORK LEVEL SCORES FOR THE DISEASE STAGE GRNS INFERRED FROM THE MOLENAAR DATASET USING NOVEL INFERENCE METHOD

Metric	**Stage 1**	**Stage 3**	**Stage 4**	**Stage 4M**
Weighted Degree Assortativity	-0.449	-0.295	-0.288	-0.284
Degree Assortativity	-0.476	-0.312	-0.300	-0.278
Diameter	0.608	1.760	1.545	1.859
Network Clustering	0.292	0.482	0.399	0.347
Largest Clique Size	84	134	105	72

TABLE 6.6 NETWORK LEVEL SCORES FOR THE NEUROBLASTOMA DISEASE STAGE GRNS INFERRED FROM THE WANG DATASET USING NOVEL INFERENCE METHOD

Metric	**Class 1**	**Class 3**	**Class 4**	**Class 4M**
Weighted Degree Assortativity	-0.276	-0.077	-0.289	-0.419
Degree Assortativity	-0.265	-0.069	-0.280	-0.424

Diameter	1.654	2.227	2.093	2.545
Network Clustering	0.328	0.546	0.330	0.455
Largest Clique Size	63	122	76	114

Looking at the network level metric scores for the two neuroblastoma datasets, some patterns emerge between metric level score and neuroblastoma category. As the neuroblastoma stage increases from stage 1 to stage 4M, the weighted degree assortativity and degree assortativity values both increase for the Molenaar dataset, although this is not the case with the Wang dataset. For both datasets, the value of the diameter is greatest in the stage 4M category networks. As noted before, diameter can be indicative of how easily genes are able to communicate in a network; a low relative value might indicate genes are able to communicate with each other much easier than when there is a high relative diameter value. As the most advanced neuroblastoma stage, stage 4M, has the highest diameter for both datasets, this could be an indication that genes involved in certain cellular processes are not able to communicate easily with each other, contributing to the more advanced stage of neuroblastoma. Between the two datasets, only the diameter values correlate perfectly, again indicating that the two datasets are somewhat different. As noted before, the most important observations to be made about the networks on a network level could possibly be made about the size and composition of the largest unique cliques in the networks. This is examined in depth in the following section.

6.3 Largest Clique Calculation and Analysis

The size and composition of the largest unique cliques in the GRNs for the different categories can be informative. It is also of interest to compare the largest unique cliques for the categories in the networks inferred from both datasets. A large number of common genes in the same category cliques across both datasets would be indicative of the same genes being involved in important cellular processes for specific neuroblastoma stages.

As well as identifying common and unique genes in these cliques, it is informative to see whether these genes have been previously identified as being neuroblastoma genes of interest and appear on the list of Génie neuroblastoma genes. The list of genes that comprise the largest unique clique in each network can also be checked for enrichment using the genesetDB website. The table below for the networks inferred from the Molenaar dataset shows the number of unique genes in each clique, whether these genes are present in the largest unique cliques in the other dataset, and also whether any of the genes in the clique appear on the Génie neuroblastoma gene list.

TABLE 6.7 NEUROBLASTOMA GENES OF INTEREST IN THE LARGEST UNIQUE CLIQUES OF THE NEUROBLASTOMA DISEASE STAGE GRNS INFERRED FROM THE MOLENAAR DATASET

	Clique Size	Genes present in other Molenaar cliques	Genes present in Wang Stage 1 clique	Genes present in Wang Stage 3 clique	Genes present in Wang Stage 4 clique	Genes present in Wang Stage 4M clique	Number of Génie genes and score
Molenaar Stage 1	84	4	0	12	1	1	13, score of 3618
Molenaar Stage 3	134	1	0	1	4	0	12, score of 3000
Molenaar Stage 4	105	5	0	26	2	0	21, score of 5211
Molenaar Stage 4M	72	6	0	15	1	0	6, score of 1484

Across the cliques, a number of neuroblastoma genes that appear the list generated by Génie are present. Stage 4 has both the highest number of Génie genes, and also the highest score of genes identified. Stage 4M has the lowest number of genes identified, and the lowest score. 52 genes that appear on the list generated by Génie for neuroblastoma are present in the cliques, which represents just over 10% of the complete list used. This is quite a significant proportion, and is suggestive that the genes in the cliques are important neuroblastoma genes. A number of genes in these cliques are also present in the cliques in the Wang dataset, most notably the largest unique clique for stage 4 that has 33 genes in cliques in the Wang dataset. However, there are only 3 genes common to the same category clique in both datasets, another indicator of how different the two datasets are. The Venn diagram in figure B.6 of the appendix shows the distribution of the genes in the largest unique cliques in the GRNs inferred from the Molenaar dataset.

There are 395 genes in total in the four cliques, with 379 being unique to one clique. There are no genes common to either all four cliques, or three cliques. However, none of the cliques are comprised entirely of unique genes. The high proportion of unique genes and very few shared genes suggests that the cellular processes are very different in the cliques in the different neuroblastoma categories. This result corresponds to the biology, that suggests that the neuroblastoma stages are different from each other and as such different cellular processes are involved in the different stages.

Table 6.8 below details the number of unique genes in each clique inferred for the networks from the Wang dataset, whether these genes are present in the largest unique cliques in the other dataset, and also whether any of the genes in the clique appear on the Génie neuroblastoma gene list.

TABLE 6.8 NEUROBLASTOMA GENES OF INTEREST IN THE LARGEST UNIQUE CLIQUES OF THE NEUROBLASTOMA DISEASE STAGE GRNS INFERRED FROM THE WANG DATASET

	Clique Size	Genes present in other Wang cliques	Genes present in Molenaar Stage 1 clique	Genes present in Molenaar Stage 3 clique	Genes present in Molenaar Stage 4 clique	Genes present in Molenaar Stage 4M clique	Number of Génie genes and score
Wang Stage 1 Clique	63	0	0	0	0	0	3, score of 525
Wang Stage 3 Clique	122	0	12	1	26	12	16, score of 4791
Wang Stage 4 Clique	76	0	1	4	2	1	4, score of 1237
Wang Stage 4M Clique	114	0	1	0	1	0	4, score of 1580

As was the case with the cliques in the networks inferred from the Molenaar dataset, across these cliques a number of neuroblastoma genes that appear the list generated by Génie are present. Stage 3 has both the highest number of Génie genes, and also the highest score of genes identified. Stage 1 has the lowest number of genes identified, and the lowest score. A substantially lower number of genes that appear on the list generated by Génie for neuroblastoma are present in the cliques, 27, compared to the Molenaar dataset, 52. This could suggest that the cliques inferred for the networks in the Molenaar dataset are more biologically relevant for neuroblastoma, although this is based on the implication that presence of genes in the Génie neuroblastoma list is indicative of biological relevance. A number of genes in these cliques are also present in the cliques in the Molenaar dataset, most notably the largest unique clique for stage 3 that has 51 genes in cliques in the Wang dataset. However, there are only 3 genes common to the same category clique in both datasets, another indicator of how different the two datasets are. The Venn diagram in figure B.7 of the appendix shows the distribution of the genes in the largest unique cliques in the GRNs inferred from the Wang dataset.

There are 375 genes in total in the four cliques. Unlike the cliques in the Molenaar dataset, there are no genes common to either two, three or four cliques, with each gene only appearing in one clique. This result is even more indicative that the cellular processes are very different in the cliques in the different neuroblastoma categories than for the Molenaar dataset. Again, this result corresponds to the biology, that suggests that the neuroblastoma stages are different from each other and as such different cellular processes are involved in the different stages.

6.4 Enrichment Results for the Largest Unique Cliques in the GRNs Inferred from the Molenaar Dataset

Having investigated the distribution of genes in the cliques in the two datasets, the next step is to investigate the enrichment of these genes using GenesetDB. In this section, the enrichment results for the genes that comprise the largest unique cliques in the GRNs inferred from the Molenaar dataset will be analysed. Table A.22 in the appendix shows the top 10 results for the genes that comprise the largest unique clique in the GRN inferred for stage 1 in the Molenaar dataset. Of interest is the presence of enrichment results related to immune function. As noted by De Preter et al in a 2006 study [128], neuroblastomas are characterised by an over-representation of genes involved in immune response, cell growth, and cell cycle.

Table A.23 in the appendix shows the top 10 results for enrichment of the genes that comprise the stage 3 largest unique clique. The 10th result detailing Interferon alpha/beta signaling is of note; a 2010 study by Dedoni et al [129] detailed how Interferon-β induces apoptosis in human SH-SY5Y neuroblastoma cells.

Table A.24 of the appendix shows the top 10 results for enrichment of the genes that comprise the stage 4 largest unique clique. Unlike the enrichment results for the two previous cliques, there are no results that immediately stand out for neuroblastoma.

The enrichment results for the genes that comprise the largest unique clique in stage4M network inferred from the Molenaar dataset are shown in table A.25 of the appendix. The 10th result, detailing negative regulation of transforming growth factor beta (TGF-β) receptor signaling pathway, is of particularly note for this clique; a 2000 study by Iolascon et al [130] specifically highlighted reduced expression of TGF-β in high stage neuroblastomas. That it appears in the top ten enrichment results for stage 4M, the highest stage neuroblastoma, corresponds with this study.

6.5 Enrichment Results for the Largest Unique Cliques in the GRNs Inferred from the Wang Dataset

In this section, the enrichment results for the genes that comprise the largest unique cliques in the GRNs inferred from the Wang dataset will be analysed. For the list of genes that comprise the largest unique clique in the GRN inferred for stage 1, only one result is returned from GenesetDB, shown in table A.26 of the appendix. This implies that this combination of genes has not been identified as being involved in biological processes.

Table A.27 of the appendix shows the top 10 results for the enrichment of the genes that comprise the stage 3 largest unique clique. The 2nd result in the table, detailing SP1 gene regulation, should be noted as a 2003 study by Tuthill et al [131] highlighted the role that SP1 plays in terms of MYCN amplification. As MYCN is not amplified in the stage 3 samples in the Wang dataset, this result would suggest that SP1 might play a role in this.

For the list of genes that comprise the largest unique clique in stage 4, no enrichment results are returned from GenesetDB. This would imply that this combination of genes has not been identified as being involved in biological processes.

The top 10 enrichment results for the genes that comprise the stage 4M largest unique clique are shown in table A.28 of the appendix. There are no results that immediately stand out for neuroblastoma.

6.6 Node Level Metrics of the GRNs Inferred from the Molenaar Dataset

Having investigated the network level metrics, and the largest unique cliques in the networks, the focus is now the node level metrics. As before, a combined metric comprised of weighted degree, weighted betweenness and weighted closeness, is used to infer a list of top 20 ranked genes for this survival category, and all subsequent categories. Enrichment and the presence of previously identified neuroblastoma genes will be investigated for the combined list. Correlations will be calculated for the metrics, and the metrics will be scored using the 500 top ranked genes for neuroblastoma returned by Génie.

6.6.1 Stage 1 Molenaar Dataset

Table 6.9 shows the top 20 ranked genes for the combined metric in the GRN for the stage 1 category inferred from the Molenaar dataset. Four of these genes in the table appear in the list of 500 neuroblastoma genes generated by Génie; AKR1B1, NGFR, NOV, and CCL2. This represents quite a high proportion of genes having been previously identified as interest for neuroblastoma. None of the genes that appear in the top 20 ranked list appear in any of the top

20 ranked for the other categories. As was the case with the largest unique cliques, this suggests clearly distinguishable differences between the neuroblastoma categories in terms of the genes that are playing important roles in cellular processes.

A number of interesting enrichment results are returned for this list of genes, shown in table A.29. As well as the previously mentioned SP1 regulation, the 8th result in relation to TNFR2 is worth highlighting as TNFR2 is proposed as a novel target to modulate cell responses to nerve growth factor by Takei et al [132], a process involved in neuroblastoma.

TABLE 6.9 TOP 20 RANKED GENES FOR COMBINED METRIC IN NEUROBLASTOMA STAGE 1 GRN INFERRED FROM MOLENAAR DATASET

Gene	***Combined Metric Rank***	***Gene***	***Combined Metric Rank***
SLC22A4	1	SMPDL3A	11
DOCK2	2	NGFR	12
MEOX2	3	FCER1A	13
CLEC10A	4	NOV	14
IKBKAP	5	SLC26A2	15
ASB4	6	SH3BP5	16
PDE4DIP	7	TNFAIP3	17
AKR1B1	8	BTG2	18
TM7SF2	9	CCL2	19
ABCB9	10	ARHGAP25	20

Table 6.10 below shows the Spearman Rank correlation scores for the rankings assigned to the genes by each metric in the stage 1 GRN inferred from the Molenaar dataset.

TABLE 6.10 SPEARMAN RANK CORRELATION SCORES OF RANKINGS ASSIGNED BY THE METRICS TO THE GENES IN STAGE 1 GRN INFERRED FROM MOLENAAR DATASET

	Degree Centrality	**Weighted Degree**	**Eigenvector Centrality**	**Closeness Centrality**	**Betweenness Centrality**
Degree Centrality	1	0.998	0.926	0.835	0.756
Weighted Degree	0.998	1	0.929	0.826	0.747
Eigenvector Centrality	0.926	0.929	1	0.773	0.670
Closeness Centrality	0.835	0.826	0.773	1	0.747
Betweenness Centrality	0.756	0.747	0.670	0.747	1

From table 6.10 above, it can be seen that all of metrics have a positive correlation greater than 0.67. These strong correlations imply that there is a substantial overlap in the genes that the metrics rank highly, as was the case for a lot of the metrics in the last chapters. Therefore, we would expect that there will be quite a lot of overlap in the genes that these metrics identify.

The list of genes generated by Génie for neuroblastoma will be used to score the metrics, limited to the top 500 ranking genes. This is due to a great deal of genes returned, and some specificity of results is desired. Starting with the top 20 ranked genes by each metric, table 6.11 below shows the scores that are obtained.

TABLE 6.11 GÉNIE SCORES AND METRIC RANKING FOR THE TOP 20 RANKED GENES FOR EACH METRIC IN STAGE 1 GRN INFERRED FROM MOLENAAR DATASET

Metric	**Number of Génie neuroblastoma genes identified**	**Génie Score**	**Metric Rank**
Combined Metric	4	1106	1st
Weighted Betweenness	3	720	2nd
Eigenvector Centrality	3	707	3rd
Weighted Closeness	3	634	4th
Degree Centrality	2	589	5th
Weighted Degree	2	589	5th

The combined metric performs best, followed by weighted betweenness and eigenvector. The combined metric identifies four genes, as noted earlier; the weighted betweenness and eigenvector centrality identify three genes. Applying the scoring system to the top 100 ranked genes for each metric, the results in table 6.12 below are given.

TABLE 6.12 GÉNIE SCORES AND METRIC RANKING FOR THE TOP 100 RANKED GENES FOR EACH METRIC IN STAGE 1 GRN INFERRED FROM MOLENAAR DATASET

Metric	**Number of Génie neuroblastoma genes identified**	**Génie Score**	**Metric Rank**
Degree Centrality	14	3924	1st
Weighted Degree	14	3924	1st
Combined Metric	14	3749	3rd
Weighted Betweenness	8	2410	4th
Weighted Closeness	11	2384	5th

Eigenvector Centrality	6	1919	6th

Applying the scoring system to the top 100 genes identified by the metrics, the weighted degree and degree centrality are the joint best performing metrics, identifying the same 14 genes. The combined metric also identifies 14 genes, but these are not as high scoring. The weighted betweenness and eigenvector centrality metrics perform relatively poorly. Finally, extending the scoring system to the top 500 genes, the following results shown in table 6.13 below are given.

Table 6.13 Génie scores and metric ranking for the top 500 ranked genes for each metric in stage 1 GRN inferred from Molenaar dataset

Metric	**Number of Génie neuroblastoma genes identified**	**Génie Score**	**Metric Rank**
Combined Metric	41	10823	1st
Degree Centrality	39	10717	2nd
Weighted Degree	39	10717	2nd
Eigenvector Centrality	37	10360	4th
Weighted Closeness	40	10032	5th
Weighted Betweenness	35	9464	6th

The combined metric performs best, followed by weighted degree and degree centrality. Four of the top 25 ranked genes by Génie for neuroblastoma are identified by these metrics; AKT1 ranked 3rd, NFKB1 ranked 14th, STAT3 ranked 17th, and PLAU ranked 21st. This result suggests that these metrics are able to identify biologically relevant genes of interest in the network inferred from the stage 1 samples in the Molenaar dataset.

Across the whole network, 142 out of the 500 neuroblastoma genes generated by Génie are present, with 41 of these identified by the combined metric. However, only 4 out of the top 50 Génie ranked neuroblastoma genes are present in the top 500 ranked neuroblastoma genes for any of the metrics. This shows that many of the genes that Génie highlights as being important for neuroblastoma are not identified by the network metrics in this category. This could be due to two reasons. Either the network inference method and metrics are not accurate in re-constructing and identifying the gene regulatory processes that take place, or these genes identified by Génie do not in fact play an important role in this neuroblastoma. It should also be noted that there is no stage specific information present with the genes identified by Génie, i.e. specific genes are not associated with specific neuroblastoma stages.

6.6.2 Stage 3 Molenaar Dataset

Table 6.14 shows the top 20 ranked genes for the combined metric in the GRN inferred for the stage 3 category from the Molenaar dataset. One of the genes in the table, , ADCYAP1, appears in the list of 500 neuroblastoma genes generated by Génie, ranked as the 330th most important gene for neuroblastoma. Again, none of the genes that appear in the top 20 ranked list appear in any of the top 20 ranked for the other categories. As was the case with the previous category, this suggests that different genes are playing important roles in the different neuroblastoma stages.

No enrichment results are returned for the list of top 20 ranked genes, implying that this particular combination of genes has not been identified as being involved in any significant biological processes. The presence of only one gene in the top 500 list returned by Génie also suggests that this list of genes might not be as biologically relevant as the previous list.

TABLE 6.14 TOP 20 RANKED GENES FOR COMBINED METRIC IN STAGE 3 GRN INFERRED FROM MOLENAAR DATASET

Gene	*Combined Metric Rank*	*Gene*	*Combined Metric Rank*
LSAMP	1	CETN2	11
CD59	2	ADCYAP1	12
GHITM	3	ATP6V1D	13
PSMD10	4	ARL1	14
SNCG	5	HIPK3	15
CRYGC	6	PEA15	16
CHUK	7	NEBL	17
ANAPC13	8	PDE8B	18
RIT2	9	TOR1AIP1	19
ZNF365	10	UBE2V2	20

Table 6.15 below shows the Spearman Rank correlation scores for the rankings assigned to the genes by each metric in the stage 3 GRN inferred from the Molenaar dataset.

TABLE 6.15 SPEARMAN RANK CORRELATION SCORES OF RANKINGS ASSIGNED BY THE METRICS TO THE GENES IN STAGE 3 GRN INFERRED FROM MOLENAAR DATASET

	Degree Centrality	**Weighted Degree**	**Eigenvector Centrality**	**Closeness Centrality**	**Betweenness Centrality**
Degree Centrality	1	0.998	0.905	0.877	0.771
Weighted Degree	0.998	1	0.902	0.870	0.765
Eigenvector Centrality	0.905	0.902	1	0.889	0.643

Closeness Centrality	0.877	0.870	0.889	1	0.734
Betweenness Centrality	0.771	0.765	0.643	0.734	1

From table 6.15 above, it can be seen that all of metrics have a positive correlation greater than 0.64. The correlation scores are very similar to the scores for the previous neuroblastoma stage. These strong correlations again imply that there is a substantial overlap in the genes that the metrics rank highly, as was the case for a lot of the metrics in the last chapters. Therefore, we would expect that there will be quite a lot of overlap in the genes that these metrics identify.

Applying the scoring system to the top 20, top 100, and top 500 genes identified by each of metrics, the metrics can be ranked. The scores for the top 20 genes identified by each metric are shown in table 6.16 below.

TABLE 6.16 GÉNIE SCORES AND METRIC RANKING FOR THE TOP 20 RANKED GENES FOR EACH METRIC IN STAGE 3 GRN INFERRED FROM MOLENAAR DATASET

Metric	**Number of Génie neuroblastoma genes identified**	**Génie Score**	**Metric Rank**
Weighted Betweenness	1	171	1st
Weighted Closeness	1	171	1st
Combined Metric	1	171	1st
Degree Centrality	0	0	4th
Weighted Degree	0	0	4th
Eigenvector Centrality	0	0	4th

As noted previously, only one gene on the top 500 ranked list is identified by the combined rank, and this is also the case for both the weighted degree and degree centrality, which are the best performing metrics along with the combined metric. Applying the scoring system to the top 100 ranked genes for each metric, the results in table 6.17 below are given.

TABLE 6.17 GÉNIE SCORES AND METRIC RANKING FOR THE TOP 100 RANKED GENES FOR EACH METRIC IN STAGE 3 GRN INFERRED FROM MOLENAAR DATASET

Metric	**Number of Génie neuroblastoma genes identified**	**Génie Score**	**Metric Rank**
Eigenvector Centrality	10	2272	1st
Degree Centrality	6	1236	2nd
Weighted Degree	6	1236	2nd
Weighted Closeness	4	752	4th
Combined Metric	5	585	5th
Weighted Betweenness	3	365	6th

Applying the scoring system to the top 100 genes identified by the metrics, the eigenvector centrality is the best performing metric, identifying ten genes. One of these genes identified, FGF2, is the 30th ranked gene for neuroblastoma. Fewer genes are identified by the metrics than for the last neuroblastoma stage. Finally, extending the scoring system to the top 500 genes, the following results shown in table 6.18 below are given.

TABLE 6.18 GÉNIE SCORES AND METRIC RANKING FOR THE TOP 500 RANKED GENES FOR EACH METRIC IN STAGE 3 GRN INFERRED FROM MOLENAAR DATASET

Metric	**Number of Génie neuroblastoma genes identified**	**Génie Score**	**Metric Rank**
Eigenvector Centrality	38	10478	1st
Degree Centrality	30	8225	2nd
Weighted Degree	30	8225	2nd
Combined Metric	26	6517	4th
Weighted Closeness	25	6509	5th
Weighted Betweenness	23	5979	6th

The eigenvector centrality out-performs the other metrics when the scoring is applied to the top 500 genes identified by each metric. However, only one of the top 25 ranked genes for neuroblastoma is identified by eigenvector centrality, compared to 4 of the top 25 by the best performing metric for the previous neuroblastoma stage. Extending this, only 3 of the top 50 ranked genes are identified, again worse than the best performing metric for the last category. All of the metrics perform worse than for the previous category. This is also shown across the whole network. 78 out of the 500 neuroblastoma genes generated by Génie are present, fewer genes identified than in the last category.

6.6.3 Stage 4 Molenaar Dataset

Table 6.19 shows the top 20 ranked genes for the combined metric in the stage 4 GRN for the Molenaar dataset. None of the genes in the table appear in the list of 500 neuroblastoma genes generated by Génie. As with the two previous categories, none of the genes that appear in the table are present in the top 20 rankings for any of the other categories.

Entering this list of genes into GenesetDB with a FDR of 0.05, 2 results are returned, shown in table A.30 of the appendix. This follows on from the previous category, where no enrichment results were returned. This suggests that checking the top ranked genes for neuroblastoma category might not be particularly informative.

TABLE 6.19 TOP 20 RANKED GENES FOR COMBINED METRIC IN STAGE 4 GRN INFERRED FROM MOLENAAR DATASET

Gene	*Combined Metric Rank*	*Gene*	*Combined Metric Rank*
TRIP12	1	ZNF337	11
MIPEP	2	CAPNS1	12
CHGA	3	WFDC2	13
SLC8A2	4	PDCD6	14
INS	5	UBE2M	15
LYPLA2	6	CDH4	16
RPS6KB1	7	MPV17	17
CELSR1	8	PAK4	18
PSMC4	9	SLC35A2	19
PNMT	10	DPF1	20

Table 6.20 below shows the Spearman Rank correlation scores for the rankings assigned to the genes by each metric in the stage 4 GRN inferred from the Molenaar dataset.

TABLE 6.20 SPEARMAN RANK CORRELATION SCORES OF RANKINGS ASSIGNED BY THE METRICS TO THE GENES IN STAGE 4 GRN INFERRED FROM MOLENAAR DATASET

	Degree Centrality	**Weighted Degree**	**Eigenvector Centrality**	**Closeness Centrality**	**Betweenness Centrality**
Degree Centrality	1	0.998	0.896	0.847	0.813
Weighted Degree	0.998	1	0.893	0.841	0.806
Eigenvector Centrality	0.896	0.893	1	0.811	0.691
Closeness Centrality	0.847	0.841	0.811	1	0.829
Betweenness Centrality	0.813	0.806	0.691	0.829	1

There are again strong correlation scores between the metrics, with the minimum correlation score being 0.69. As before, this implies significant overlap in the genes ranked highly by the different metrics. Applying the scoring system to the top 20, top 100, and top 500 genes identified by each of metrics, the metrics can be ranked. The scores for the top 20 genes identified by each metric are shown in table 6.21 below.

TABLE 6.21 GÉNIE SCORES AND METRIC RANKING FOR THE TOP 20 RANKED GENES FOR EACH METRIC IN STAGE 4 GRN INFERRED FROM MOLENAAR DATASET

Metric	**Number of Génie neuroblastoma genes identified**	**Génie Score**	**Metric Rank**
Eigenvector Centrality	6	1479	1st
Weighted Degree	5	1276	2nd
Degree Centrality	3	676	3rd

Weighted Betweenness	1	461	4th
Weighted Closeness	0	0	5th
Combined Metric	0	0	5th

The eigenvector and weighted degree centrality metrics perform strongly, identifying six and five genes, respectively, from the top 500 ranked neuroblastoma genes. Both these metrics identify the 8th ranked gene MMP2 in their respective top 20 rankings. The weighted closeness and combined metric do not identify any genes. Applying the scoring system to the top 100 ranked genes for each metric, the results in table 6.22 below are given.

TABLE 6.22 GÉNIE SCORES AND METRIC RANKING FOR THE TOP 100 RANKED GENES FOR EACH METRIC IN STAGE 4 GRN INFERRED FROM MOLENAAR DATASET

Metric	**Number of Génie neuroblastoma genes identified**	**Génie Score**	**Metric Rank**
Eigenvector Centrality	22	5676	1st
Degree Centrality	16	4557	2nd
Weighted Degree	17	4469	3rd
Combined Metric	10	3124	4th
Weighted Betweenness	9	2845	5th
Weighted Closeness	5	1625	6th

The eigenvector centrality again is the best performing metric, followed by the degree and weighted degree centrality metrics. Four of the top 50 ranked genes for neuroblastoma; MMP2, MMP14, ADAM17, and MUC1, are identified by eigenvector centrality from the 22 genes it identifies. The metrics perform better than in the previous two neuroblastoma categories. Finally, extending the scoring system to the top 500 genes, the following results shown in table 6.23 below are given.

TABLE 6.23 GÉNIE SCORES AND METRIC RANKING FOR THE TOP 500 RANKED GENES FOR EACH METRIC IN STAGE 4 GRN INFERRED FROM MOLENAAR DATASET

Metric	**Number of Génie neuroblastoma genes identified**	**Génie Score**	**Metric Rank**
Eigenvector Centrality	64	17763	1st
Degree Centrality	54	15091	2nd
Weighted Degree	54	15091	2nd
Weighted Closeness	50	14366	4th
Combined Metric	48	13480	5th
Weighted Betweenness	47	13380	6th

Overall, the metrics perform better for this category than for the two previous Molenaar neuroblastoma categories, identifying a significant number of top 500 ranked neuroblastoma genes. The eigenvector centrality again is the best performing metric, 9 of the 64 ranked genes it identifies are in the top 50. The degree and weighted degree metrics are the next best performing metrics, each identifying the same 54 neuroblastoma genes. Even the worst performing metric, weighted betweenness, for this neuroblastoma identifies 47 genes and performs better than the best performing metrics for the two previous neuroblastoma categories. These results suggest that for this neuroblastoma category, the metrics perform much better at identifying neuroblastoma genes of interest. The stronger performance of the metrics is reinforced by the identification of 156 out of the 500 neuroblastoma genes generated by Génie across the whole network, more than for the two previous neuroblastoma categories.

6.6.4 Stage 4M Molenaar Dataset

Table 6.24 shows the top 20 ranked genes for the combined in the stage 4M GRN for the Molenaar dataset. Four of the genes in the table appear in the list of 500 neuroblastoma genes generated by Génie; MYCN ranked 1st, SMAD3 ranked 197th, PDGFRA ranked 223rd, and PFKFB3 ranked 383rd. Bearing in mind that the criteria for the stage 4M neuroblastoma category is MYCN amplification, this result corresponds with the underlying biology. As before, none of the genes that appear in the top 20 ranked list appear in any of the top 20 ranked for the other categories.

Looking at the enrichment results returned for this list of genes, shown in table A.31 of the appendix, the 9th result is most if interest. The genes MYCN, FES, and PDGFRA, are identified by the Cancer Genes database [133]. Although it should be noted that this is a database for identifying genes that are generally of interest for cancer genomics, and is not specific to neuroblastoma.

TABLE 6.24 TOP 20 RANKED GENES FOR COMBINED METRIC IN STAGE 4M GRN INFERRED FROM MOLENAAR DATASET

Gene	***Combined Metric Rank***	***Gene***	***Combined Metric Rank***
MYCN	1	INPP4B	11
FES	2	PFAS	12
TGIF2	3	RAB32	13
GPR125	4	HDAC9	14

SOCS2	5	STC2	15
HS3ST1	6	RSL1D1	16
SMAD3	7	PFKFB3	17
FBL	8	NFIX	18
PDGFRA	9	MGP	19
DDX10	10	APEX1	20

Table 6.25 below shows the Spearman Rank correlation scores for the rankings assigned to the genes by each metric in the stage 4M GRN inferred from the Molenaar dataset.

TABLE 6.25 SPEARMAN RANK CORRELATION SCORES OF RANKINGS ASSIGNED BY THE DIFFERENT METRICS TO THE GENES IN STAGE 4M GRN INFERRED FROM MOLENAAR DATASET

	Degree Centrality	**Weighted Degree**	**Eigenvector Centrality**	**Closeness Centrality**	**Betweenness Centrality**
Degree Centrality	1	0.99818175	0.960905689	0.834491125	0.871084923
Weighted Degree	0.99818175	1	0.95895602	0.827033544	0.868246578
Eigenvector Centrality	0.960905689	0.95895602	1	0.899516798	0.822331147
Closeness Centrality	0.834491125	0.827033544	0.899516798	1	0.775013763
Betweenness Centrality	0.871084923	0.868246578	0.822331147	0.775013763	1

The correlation scores between the metrics are again strong, with the weakest correlation having a score of 0.77. Again there will still be substantial overlap between the genes identified by the metrics.

Applying the scoring system to the top 20, top 100, and top 500 genes identified by each of metrics, the metrics can be ranked. The scores for the top 20 genes identified by each metric are shown in table 6.26 below.

TABLE 6.26 GÉNIE SCORES AND METRIC RANKING FOR THE TOP 20 RANKED GENES FOR EACH METRIC IN STAGE 4M GRN INFERRED FROM MOLENAAR DATASET

Metric	**Number of Génie neuroblastoma genes identified**	**Génie Score**	**Metric Rank**
Weighted Betweenness	4	1200	1st
Combined Metric	4	1200	1st
Eigenvector Centrality	3	1125	3rd
Degree Centrality	3	1082	4th
Weighted Degree	3	1082	4th
Weighted Closeness	3	1082	4th

The weighted betweenness is the best performing metric. In addition to identifying the genes MYCN, PDGFRA, and SMAD3, that the other metrics all identify, it also identifies PFKFB3 ranked 383rd by Génie. Applying the scoring system to the top 100 ranked genes for each metric, the results in table 6.27 below are given.

TABLE 6.27 GÉNIE SCORES AND METRIC RANKING FOR THE TOP 100 RANKED GENES FOR EACH METRIC IN STAGE 4M GRN INFERRED FROM MOLENAAR DATASET

Metric	**Number of Génie neuroblastoma genes identified**	**Génie Score**	**Metric Rank**
Weighted Closeness	12	2828	1st

Weighted Betweenness	8	2645	2nd
Degree Centrality	9	2535	3rd
Weighted Degree	9	2425	4th
Eigenvector Centrality	10	2220	5th
Combined Metric	8	2159	6th

The weighted closeness is the best performing metric, identifying 12 neuroblastoma Génie genes. TNC and PTK2B identified before, and CTNNB1 and PEG3. The weighted degree and eigenvector centrality metrics perform badly, only identifying the same gene that they identified before in their top 20 rankings. Finally, extending the scoring system to the top 500 genes, the following results shown in table 6.28 below are given.

TABLE 6.28 GÉNIE SCORES AND METRIC RANKING FOR THE TOP 500 RANKED GENES FOR EACH METRIC IN STAGE 4M GRN INFERRED FROM MOLENAAR DATASET

Metric	**Number of Génie neuroblastoma genes identified**	**Génie Score**	**Metric Rank**
Eigenvector Centrality	52	14837	1st
Weighted Degree	48	13465	2nd
Degree Centrality	47	13104	3rd
Combined Metric	40	11011	4th
Weighted Betweenness	41	10956	5th
Weighted Closeness	38	10316	6th

Applying the scoring to the top 500 genes identified by each metric, eigenvector centrality is the best performing metric, followed by the weighted degree centrality. Whilst all of the metrics perform slightly worse at identifying genes than in the previous category, they all identify 3 of the top 20 ranked genes by Génie for neuroblastoma. MYCN ranked 1st, TGFB1 ranked 7th, and HIF1A ranked 19th, are amongst the respective top 500 genes ranked by all the metrics. Although not as good a result as for the previous neuroblastoma category, this still implies that these metrics are able to identify a number of biologically significant genes.

Across the whole network, 150 out of the 500 neuroblastoma genes generated by Génie are present, a similar number to the last category. This again suggest that there are a number of genes previously identified as being important for neuroblastoma present in the network for this category, and that these genes are involved in cellular processes in the network for this category.

6.7 Metric Performance GRNs Inferred from Molenaar Dataset

Previously, each metric was ranked based on the top 20, top 100 and top 500 genes they identified from each neuroblastoma disease stage GRN for the Molenaar dataset. This ranking is based on the score of the identified genes in the Génie neuroblastoma gene list.

The results of these metric rankings can now be presented together, allowing an analysis of which metric is the best performing overall. Assigning an equal weighting to the scores of the top 20, top 100, and top 500 ranked genes for each metric in each category, the following ranks shown in table 6.29 below are given.

TABLE 6.29 OVERALL METRIC RANKS IN THE DIFFERENT NEUROBLASTOMA DISEASE STAGE GRNS INFERRED USING NOVEL ABSOLUTE Z SCORE NETWORK INFERENCE METHOD FROM THE MOLENAAR DATASET

	Stage 1 Rank	**Stage 3 Rank**	**Stage 4 Rank**	**Stage 4M Rank**	**Ranking Totals**	**Overall Rank**
Degree Centrality	2nd	2nd	2nd	3rd	9	1st
Weighted Degree	2nd	2nd	2nd	3rd	9	1st
Eigenvector Centrality	5th	1st	1st	2nd	9	1st
Combined Metric	1st	4th	4th	5th	14	4th
Weighted Betweenness	4th	6th	5th	1st	16	5th
Weighted Closeness	6th	4th	5th	5th	20	6th

Weighted degree, degree, and eigenvector centrality are the joint best performing metrics overall. The number of connections a node has in a network is the most often applied property for determining its importance in a network, so it should perhaps not be surprising to see the degree metrics performing well. If the rankings for just the top 20 genes in each category are taken into consideration, the following scores in table 6.30 are obtained.

TABLE 6.30 OVERALL METRIC RANKS IN THE DIFFERENT NEUROBLASTOMA DISEASE STAGE GRNS INFERRED USING NOVEL ABSOLUTE Z SCORE NETWORK INFERENCE METHOD FROM THE MOLENAAR DATASET BASED ON TOP 20 GENES IDENTIFIED BY EACH METRIC

	Stage 1 Rank	**Stage 3 Rank**	**Stage 4 Rank**	**Stage 4M Rank**	**Ranking Totals**	**Overall Rank**
Weighted Betweenness	2nd	1st	4th	1st	8	1st
Combined Metric	1st	1st	5th	1st	8	1st
Eigenvector Centrality	3rd	4th	1st	3rd	11	3rd
Weighted Closeness	4th	1st	5th	4th	14	4th
Weighted Degree	5th	4th	2nd	4th	15	5th
Degree Centrality	5th	4th	3rd	4th	16	6th

Weighted betweenness and the combined metric are the joint best performing metrics. The degree based metrics perform poorly this time, perhaps surprising considering the phenomenon of scale-free network where a few number of nodes have a high degree, and a large number of nodes have a low degree. If we take the ranking for the top 100 genes into account, the following results shown in table 6.31 are obtained.

TABLE 6.31 OVERALL METRIC RANKS IN THE DIFFERENT NEUROBLASTOMA DISEASE STAGE GRNS INFERRED USING NOVEL ABSOLUTE Z SCORE NETWORK INFERENCE METHOD FROM THE MOLENAAR DATASET BASED ON TOP 100 GENES IDENTIFIED

	Stage 1 Rank	**Stage 3 Rank**	**Stage 4 Rank**	**Stage 4M Rank**	**Ranking Totals**	**Overall Rank**
Degree Centrality	1st	2nd	2nd	3rd	8	1st
Weighted Degree	1st	2nd	3rd	4th	10	2nd
Eigenvector Centrality	6th	1st	1st	5th	13	3rd
Weighted Closeness	5th	4th	6th	1st	16	4th
Weighted Betweenness	4th	6th	5th	2nd	17	5th
Combined Metric	3rd	5th	4th	6th	18	6th

Degree centrality is the best performing metric again, followed by weighted degree centrality and the combined metric. Finally, taking the top 500 genes into account, the following results shown in table 6.32 are obtained.

TABLE 6.32 OVERALL METRIC RANKS IN THE DIFFERENT NEUROBLASTOMA DISEASE STAGE GRNS INFERRED USING NOVEL ABSOLUTE Z SCORE NETWORK INFERENCE METHOD FROM THE MOLENAAR DATASET BASED ON TOP 500 GENES IDENTIFIED

	Stage 1 Rank	**Stage 3 Rank**	**Stage 4 Rank**	**Stage 4M Rank**	**Ranking Totals**	**Overall Rank**
Eigenvector Centrality	4th	1st	1st	1st	7	1st
Weighted Degree	2nd	2nd	2nd	2nd	8	2nd

Degree Centrality	2nd	2nd	2nd	3rd	9	3rd
Combined Metric	1st	4th	5th	4th	14	4th
Weighted Closeness	5th	5th	4th	6th	20	5th
Weighted Betweenness	6th	6th	6th	5th	23	6th

This time eigenvector centrality just out-performs weighted degree and degree centrality, with these three metrics showing similar performance. The combined metric is the 4th best performing metric. Weighted closeness centrality, the 5th ranked metric, and weighted betweenness centrality, the 6th ranked metric, perform poorly compared to the other metrics.

6.8 Distribution of Genes in Combined Metric Top 20 Rankings List across the GRNs Inferred from Molenaar Dataset

As noted in the category sections, none of the genes that appear in the top 20 ranked genes for the combined metric for any category appear in any other. This can be seen on the Venn diagram in figure B.8 of the appendix showing the distribution of genes in the top 20 combined metric rankings list in the categories. Furthermore, extending this to look at the top 100 combined metric rankings lists, shown in figure B.9 of the appendix, this is also the case and suggests that this approach is suited to identifying specific genes associated with the different survival time

6.9 Node LevelMetrics Wang Dataset

Having investigated the network level metrics, and the largest unique cliques in the networks, the focus is now the node level metrics. As before, a combined metric comprised of weighted degree, weighted betweenness and weighted closeness, is used to infer a list of top 20 ranked genes for this survival category, and all subsequent categories. Enrichment and the presence of previously identified neuroblastoma genes will be investigated for the combined list. Correlations will be calculated for the metrics, and the metrics will be scored using the 500 top ranked genes for neuroblastoma returned by Génie.

6.9.1 Stage 1 Wang Dataset

Table 6.33 shows the top 20 ranked genes for the combined metric in the GRN inferred for neuroblastoma stage 1 in the Wang dataset. One of the genes, PIK3R1, in the table appears in the list of 500 neuroblastoma genes generated by Génie, ranked as the 327th most important gene for neuroblastoma. None of the genes that appear in the top 20 ranked list appear in any of the top 20 ranked for the other categories. As was the case with all the categories in the Molenaar dataset, this suggests that different genes are playing important roles in the different neuroblastoma stages.

Entering this list of genes into GenesetDB, only six results are returned. This implies that this combination of genes has not been identified as being involved in significant biological processes, and could suggest that this is a novel finding. The six results returned are presented in table A.32 of the appendix.

TABLE 6.33 TOP 20 RANKED GENES FOR COMBINED METRIC IN STAGE 1 GRN INFERRED FROM WANG DATASET

Gene	*Combined Metric Rank*	*Gene*	*Combined Metric Rank*
MAP3K4	1	RAP2A	11

HMX1	2	SOX4	12
ERCC5	3	ADCY7	13
UBE2J1	4	SEC63	14
PHF3	5	RAB5B	15
UPF3A	6	CLASP1	16
PIK3R1	7	DDX17	17
APC	8	CAMTA1	18
CNR1	9	MCFD2	19
ELAVL2	10	TNNT2	20

Table 6.34 below shows the Spearman Rank correlation scores for the rankings assigned to the genes by each metric in the stage 1 GRN inferred from the Wang dataset.

TABLE 6.34 SPEARMAN RANK CORRELATION SCORES OF RANKINGS ASSIGNED BY THE METRICS TO THE GENES IN STAGE 1 GRN INFERRED FROM WANG DATASET

	Degree Centrality	**Weighted Degree**	**Eigenvector Centrality**	**Closeness Centrality**	**Betweenness Centrality**
Degree Centrality	1	0.998	0.887	0.866	0.860
Weighted Degree	0.998	1	0.880	0.857	0.856
Eigenvector Centrality	0.887	0.880	1	0.965	0.769
Closeness Centrality	0.866	0.857	0.965	1	0.811
Betweenness Centrality	0.860	0.856	0.769	0.811	1

The correlation scores between the metrics are again strong, with the weakest correlation having a score of 0.76. Again there will still be substantial overlap between the genes identified by the metrics.

Applying the scoring system to the top 20, top 100, and top 500 genes identified by each of metrics, the metrics can be ranked. The scores for the top 20 genes identified by each metric are shown in table 6.35 below.

TABLE 6.35 GÉNIE SCORES AND METRIC RANKING FOR THE TOP 20 RANKED GENES FOR EACH METRIC IN STAGE 1 GRN INFERRED FROM WANG DATASET

Metric	**Number of Génie neuroblastoma genes identified**	**Génie Score**	**Metric Rank**
Degree Centrality	1	174	1st
Weighted Betweenness	1	174	1st
Weighted Closeness	1	174	1st
Combined Metric	1	174	1st
Weighted Degree	0	0	5th
Eigenvector Centrality	0	0	5th

Degree, weighted betweenness, weighted closeness, and the combined metric all identify the same neuroblastoma genes of interest, PIK3R1, on the list generated by Génie. The other metrics do not identify any genes on this list. Applying the scoring system to the top 100 ranked genes for each metric, the results in table 6.36 below are given.

TABLE 6.36 GÉNIE SCORES AND METRIC RANKING FOR THE TOP 100 RANKED GENES FOR EACH METRIC IN STAGE 1 GRN INFERRED FROM WANG DATASET

Metric	**Number of Génie neuroblastoma genes identified**	**Génie Score**	**Metric Rank**
Weighted Degree	5	1224	1st
Degree Centrality	4	1083	2nd
Eigenvector Centrality	7	1022	3rd
Combined Metric	4	797	4th
Weighted Closeness	4	775	5th
Weighted Betweenness	1	174	6th

Applying the scoring system to the top 100 genes identified by the metrics, weighted degree centrality is the best performing metric, followed by degree centrality. Although eigenvector centrality identifies the most genes, weighted degree and degree centrality identify genes with a higher score. Finally, extending the scoring system to the top 500 genes, the following results shown in table 6.37 below are given.

TABLE 6.37 GÉNIE SCORES AND METRIC RANKING FOR THE TOP 500 RANKED GENES FOR EACH METRIC IN STAGE 1 GRN INFERRED FROM WANG DATASET

Metric	**Number of Génie neuroblastoma genes identified**	**Génie Score**	**Metric Rank**
Weighted Degree	31	7284	1st
Combined Metric	28	7241	2nd
Weighted Betweenness	28	7209	3rd
Degree Centrality	30	7130	4th

Weighted Closeness	25	6855	5th
Eigenvector Centrality	25	5999	6th

Applying the scoring to the top 500 genes identified by each metric, weighted degree centrality is the best performing metric, followed by the combined metric. The metrics perform poorly at identifying top 20 ranked genes by Génie for neuroblastoma. Only weighted closeness and weighted betweenness identify one gene, TNF ranked 15th, in their respective top 500 ranked genes. These results imply that the metrics are not able to identify many biologically significant genes for this category of network.

Across the whole network, 134 out of the 500 neuroblastoma genes generated by Génie are present, compared to 142 in the stage 1 network for the Molenaar dataset. As was the case with the stage 1 network in the Molenaar dataset, only 4 out of the top 50 Génie ranked neuroblastoma genes are present in the top 500 ranked neuroblastoma genes for any of the metrics. This shows that many of the genes that Génie highlights as being important for neuroblastoma are not identified by the network metrics in this category. This is again suggestive that genes Génie identifies as being important for neuroblastoma might not be involved in cellular processes in this neuroblastoma stage, due to the similar results for the same stage in the Molenaar dataset. This is something that biologically could be the case, as it is more likely that the genes Génie identifies as being important are involved in higher stage neuroblastomas.

6.9.2 Stage 3 Wang Dataset

Table 6.38 shows the top 20 ranked genes for the combined metric in the GRN inferred for neuroblastoma stage 3 in the Wang dataset. Two genes, LGALS1 ranked 406th, and ITGA5 ranked 155th, appear in the list of 500 neuroblastoma genes generated by Génie. As before, none of the genes that appear in the top 20 ranked list for the combined metric appear in any of the top 20 ranked list for the combined metric for the other neuroblastoma disease stage categories. As was the case with the previous category, this suggests that different genes are playing important roles in the different neuroblastoma stages.

No enrichment results are returned for the list of top 20 ranked genes, implying that this particular combination of genes has not been identified as being involved in any significant biological processes. This was also the case with the list of top ranked genes for stage 3 of the Molenaar dataset.

TABLE 6.38 TOP 20 RANKED GENES FOR COMBINED METRIC IN STAGE 3 GRN INFERRED FROM WANG DATASET

Gene	***Combined Metric Rank***	***Gene***	***Combined Metric Rank***
PON2	1	PPM1F	11
GNG11	2	TIE1	12
MVP	3	MITF	13
TBC1D22A	4	PDLIM7	14
LGALS1	5	FLNA	15
ARPC1B	6	GFPT2	16
TPP1	7	LRP10	17
POLD4	8	IFITM2	18
CIB1	9	NPTX1	19
ITGA5	10	GPC5	20

Table 6.39 below shows the Spearman Rank correlation scores for the rankings assigned to the genes by each metric in the stage 3 GRN inferred from the Wang dataset.

TABLE 6.39 SPEARMAN RANK CORRELATION SCORES OF RANKINGS ASSIGNED BY THE DIFFERENT METRICS TO THE GENES IN STAGE 3 GRN INFERRED FROM WANG DATASET

	Degree Centrality	**Weighted Degree**	**Eigenvector Centrality**	**Closeness Centrality**	**Betweenness Centrality**
Degree Centrality	1	0.997598864	0.921212167	0.919433584	0.831541994
Weighted Degree	0.997598864	1	0.912415326	0.907864861	0.830431065
Eigenvector Centrality	0.921212167	0.912415326	1	0.924316126	0.693070954
Closeness Centrality	0.919433584	0.907864861	0.924316126	1	0.773817133
Betweenness Centrality	0.831541994	0.830431065	0.693070954	0.773817133	1

The correlation scores between the metrics are again strong, with the weakest correlation having a score of 0.69. Again there will still be substantial overlap between the genes identified by the metrics.

Applying the scoring system to the top 20, top 100, and top 500 genes identified by each of metrics, the metrics can be ranked. The scores for the top 20 genes identified by each metric are shown in table 6.40 below.

TABLE 6.40 GÉNIE SCORES AND METRIC RANKING FOR THE TOP 20 RANKED GENES FOR EACH METRIC IN STAGE 3 GRN INFERRED FROM WANG DATASET

Metric	**Number of Génie neuroblastoma genes identified**	**Génie Score**	**Metric Rank**
Degree Centrality	4	1415	1st
Weighted Degree	4	1415	1st
Eigenvector Centrality	4	1395	3rd
Weighted Closeness	4	1281	4th
Combined Metric	2	441	5th
Weighted Betweenness	1	95	6th

Degree, weighted degree, and weighted closeness all identify four genes; however degree and weighted degree are the best performing due to indentifying higher scoring genes. Applying the scoring system to the top 100 ranked genes for each metric, the results in table 6.41 below are given.

TABLE 6.41 GÉNIE SCORES AND METRIC RANKING FOR THE TOP 100 RANKED GENES FOR EACH METRIC IN STAGE 3 GRN INFERRED FROM WANG DATASET

Metric	**Number of Génie neuroblastoma genes identified**	**Génie Score**	**Metric Rank**
Combined Metric	14	4310	1st
Degree Centrality	11	3572	2nd
Eigenvector Centrality	12	3571	3rd
Weighted Degree	12	3553	4th
Weighted Closeness	12	3506	5th

Weighted Betweenness	8	2157	6th

Applying the scoring system to the top 100 genes identified by the metrics, the combined metric is the best performing metric, identifying 14 genes. Four of the other metrics each identify 12 genes. The performance of the weighted betweenness is noticeably worse than the other metrics, only identifying 8 genes. It is worth noting that the 7th ranked gene for neuroblastoma, TGFB1, is identified by all of the metrics apart from weighted betweenness in their respective top 100 rankings. Finally, extending the scoring system to the top 500 genes, the following results shown in table 6.42 below are given.

TABLE 6.42 GÉNIE SCORES AND METRIC RANKING FOR THE TOP 500 RANKED GENES FOR EACH METRIC IN STAGE 3 GRN INFERRED FROM WANG DATASET

Metric	**Number of Génie neuroblastoma genes identified**	**Génie Score**	**Metric Rank**
Eigenvector Centrality	68	18595	1st
Degree Centrality	68	18078	2nd
Weighted Degree	68	18078	2nd
Combined Metric	60	16522	4th
Weighted Closeness	59	16222	5th
Weighted Betweenness	48	13218	6th

The eigenvector centrality out-performs the other metrics when the scoring is applied to the top 500 genes identified by each metric. 8 of the top 50 ranked genes by Génie for neuroblastoma are amongst the top 500 genes ranked by weighted betweenness centrality. This is quite a significant result in terms of the biology, and implies that weighted betweenness centrality is a good metric to use for identifying biologically significant results. All of the metrics identify substantially more genes in this stage than for the equivalent stage in the Molenaar dataset. Eigenvector is also the best performing metric for stage 3 in the Molenaar dataset identifying 38 genes, compared to 68 genes this time. Across the whole network, 140 out of the 500 neuroblastoma genes generated by Génie are present. This compares to 78 present in the stage 3 network in the Molenaar dataset.

6.9.3 Stage 4 Wang Dataset

Table 6.43 shows the top 20 ranked genes for the combined metric in the GRN inferred for neuroblastoma stage 4 in the Wang dataset. One of the genes in the table appears in the list of 500 neuroblastoma genes generated by Génie, AKT1 ranked as the 3rd most important gene for neuroblastoma. As with the two previous categories, none of the genes that appear in the table are present in the top 20 combined metric rankings for any of the GRNs for the other neuroblastoma disease stage categories.

Table A.33 in the appendix shows the top ten enrichment results for these genes from GenesetDB. The 2nd result is worth noting; the role of TNF-α in neuroblastoma apoptosis has been highlighted in a 2011 study by Álvarez et al [134].

TABLE 6.43 TOP 20 RANKED GENES FOR COMBINED METRIC IN STAGE 4 GRN INFERRED FROM WANG DATASET

Gene	*Combined Metric Rank*	*Gene*	*Combined Metric Rank*
MTERF	1	PPP1CA	11
PPFIA1	2	IRS1	12
ADIPOR2	3	CDH4	13

POLR2G	4	AKT1	14
AP1G2	5	AMOT	15
TSNAX	6	SF3B2	16
IDH3B	7	BNIP1	17
PVR	8	PPP6C	18
COX8A	9	FBXW11	19
ZNF410	10	GLT8D1	20

Table 6.44 below shows the Spearman Rank correlation scores for the rankings assigned to the genes by each metric in the stage 4 GRN inferred from the Wang dataset.

TABLE 6.44 SPEARMAN RANK CORRELATION SCORES OF RANKINGS ASSIGNED BY THE DIFFERENT METRICS TO THE GENES IN STAGE 4 GRN INFERRED FROM WANG DATASET

	Degree Centrality	**Weighted Degree**	**Eigenvector Centrality**	**Closeness Centrality**	**Betweenness Centrality**
Degree Centrality	1	0.998	0.940	0.866	0.869
Weighted Degree	0.998	1	0.936	0.858	0.865
Eigenvector Centrality	0.940	0.936	1	0.922	0.803
Closeness Centrality	0.866	0.858	0.922	1	0.835
Betweenness Centrality	0.869	0.865	0.803	0.835	1

The correlation scores between the metrics are very strong, with the weakest correlation having a score of 0.8. As a result, there will be substantial overlap between the genes identified by the metrics.

Applying the scoring system to the top 20, top 100, and top 500 genes identified by each of metrics, the metrics can be ranked. The scores for the top 20 genes identified by each metric are shown in table 6.45 below.

TABLE 6.45 GÉNIE SCORES AND METRIC RANKING FOR THE TOP 20 RANKED GENES FOR EACH METRIC IN STAGE 4 GRN INFERRED FROM WANG DATASET

Metric	**Number of Génie neuroblastoma genes identified**	**Génie Score**	**Metric Rank**
Weighted Closeness	2	768	1st
Degree Centrality	1	498	2nd
Weighted Degree	1	498	2nd
Weighted Betweenness	1	498	2nd
Combined Metric	1	498	2nd
Eigenvector Centrality	0	0	6th

Weighted closeness is the best performing metric, identifying two genes. As well as the previously mentioned AKT1, the 231st ranked gene for neuroblastoma ENPP2 is also identified. Degree, weighted degree, weighted betweenness and the combined metric all identify AKT1. Applying the scoring system to the top 100 ranked genes for each metric, the results in table 6.46 below are given.

TABLE 6.46 GÉNIE SCORES AND METRIC RANKING FOR THE TOP 100 RANKED GENES FOR EACH METRIC IN STAGE 4 GRN INFERRED FROM WANG DATASET

Metric	**Number of Génie neuroblastoma genes identified**	**Génie Score**	**Metric Rank**
Degree Centrality	8	2422	1st
Weighted Degree	7	2166	2nd
Weighted Betweenness	6	2098	3rd
Eigenvector Centrality	7	2047	4th
Combined Metric	5	1553	5th
Weighted Closeness	3	1120	6th

Applying the scoring system to the top 100 genes identified by the metrics, degree centrality is the best performing metric, identifying 8 genes. The weighted closeness centrality performs badly compared to the other metrics, only identifying 3 genes. Finally, extending the scoring system to the top 500 genes, the following results shown in table 6.47 below are given.

TABLE 6.47 GÉNIE SCORES AND METRIC RANKING FOR THE TOP 500 RANKED GENES FOR EACH METRIC IN STAGE 4 GRN INFERRED FROM WANG DATASET

Metric	**Number of Génie neuroblastoma genes identified**	**Génie Score**	**Metric Rank**
Weighted Betweenness	29	9047	1st
Degree Centrality	25	8196	2nd
Weighted Degree	24	7804	3rd
Eigenvector Centrality	24	7451	4th
Combined Metric	25	7434	5th

Weighted Closeness	22	6440	6th

The weighted betweenness is the best performing metric, followed by the degree centrality. Compared to the stage 4 category of the Molenaar dataset, the metrics identify substantially fewer genes. The best performing metric for the Molenaar stage 4 dataset identifies 9 out of the top ranked genes by Génie for neuroblastoma; here the weighted betweenness only identifies 4. Across the whole network, 149 out of the 500 neuroblastoma genes generated by Génie are present, representing a similar figure to the 156 present in the Molenaar stage 4 network.

6.9.4 Stage 4M Wang Dataset

Table 6.48 shows the top 20 ranked genes for the combined metric in the GRN inferred for neuroblastoma stage 4M in the Wang dataset. As with the previous category, one of the genes in the table appears in the list of 500 neuroblastoma genes generated by Génie. Additionally, as was the case with the previous category, this gene is very highly ranked by Génie. This time the top ranked gene for neuroblastoma, MYCN, is present. As with all the categories for both of the datasets, none of the genes that appear in the table are present in the top 20 combined metric rankings for any of the other neuroblastoma disease stage categories.

Looking at the enrichment results for this list of genes shown in table A.34 of the appendix, the presence of results detailing purine nucleotide metabolism is of note. In a 1985 study, Kaplinsky et al [135] noted that purine pathway activity was 2-3 times higher in metastatic neuroblastoma cell lines. Metastasis is associated with the final stages of cancer, so it is biologically significant that this enrichment result appears in this neuroblastoma category.

TABLE 6.48 TOP 20 RANKED GENES FOR COMBINED METRIC IN STAGE 4M GRN INFERRED FROM WANG DATASET

Gene	*Combined Metric Rank*	*Gene*	*Combined Metric Rank*
MYCN	1	RSL1D1	11
PAICS	2	RPL14	12
DDX1	3	TFAP4	13
HSPD1	4	FBL	14
TGIF2	5	GMPS	15
NPM1	6	GPR125	16
EXOSC7	7	BAZ1A	17
TRAP1	8	PRDX6	18
TKT	9	PFAS	19
IMPDH2	10	DDX10	20

Table 6.49 below shows the Spearman Rank correlation scores for the rankings assigned to the genes by each metric in the stage 4M GRN inferred from the Wang dataset.

TABLE 6.49 SPEARMAN RANK CORRELATION SCORES OF RANKINGS ASSIGNED BY THE METRICS TO THE GENES IN STAGE 4M GRN INFERRED FROM WANG DATASET

	Degree Centrality	**Weighted Degree**	**Eigenvector Centrality**	**Closeness Centrality**	**Betweenness Centrality**
Degree Centrality	1	0.996	0.979	0.936	0.842
Weighted Degree	0.996	1	0.975	0.929	0.840

Eigenvector Centrality	0.979	0.975	1	0.959	0.801
Closeness Centrality	0.936	0.929	0.959	1	0.779
Betweenness Centrality	0.842	0.840	0.801	0.779	1

The correlation scores between the metrics are strong, with the weakest correlation having a score of 0.78. As a result, there will be substantial overlap between the genes identified by the metrics.

Applying the scoring system to the top 20, top 100, and top 500 genes identified by each of metrics, the metrics can be ranked. The scores for the top 20 genes identified by each metric are shown in table 6.50 below.

Table 6.50 Génie scores and metric ranking for the top 20 ranked genes for each metric in stage 4M GRN inferred from Wang dataset

Metric	**Number of Génie neuroblastoma genes identified**	**Génie Score**	**Metric Rank**
Weighted Betweenness	3	1178	1st
Degree Centrality	1	500	2nd
Weighted Degree	1	500	2nd
Weighted Closeness	1	500	2nd
Eigenvector Centrality	1	500	2nd
Combined Metric	1	500	2nd

The weighted betweenness is the best performing metric, identifying three genes; MYCN ranked 1st, ODC1 ranked 301st, and HGF ranked 23rd. All of the other metrics identify MYCN. Applying the scoring system to the top 100 ranked genes for each metric, the results in table 6.51 below are given.

TABLE 6.51 GÉNIE SCORES AND METRIC RANKING FOR THE TOP 100 RANKED GENES FOR EACH METRIC IN STAGE 4M GRN INFERRED FROM WANG DATASET

Metric	**Number of Génie neuroblastoma genes identified**	**Génie Score**	**Metric Rank**
Weighted Betweenness	3	1178	1st
Degree Centrality	2	953	2nd
Weighted Degree	2	953	2nd
Weighted Closeness	2	953	2nd
Eigenvector Centrality	2	953	2nd
Combined Metric	2	953	2nd

Again the weighted betweenness is the best performing metric, but does not identify any additional genes. All of the other metrics identify one more gene this time, NME1, ranked 48th. The performance of the metric is relatively poor, as can be seen by the very small number of genes that they identify. Finally, extending the scoring system to the top 500 genes, the following results shown in table 6.52 below are given.

TABLE 6.52 GÉNIE SCORES AND METRIC RANKING FOR THE TOP 500 RANKED GENES FOR EACH METRIC IN STAGE 4M GRN INFERRED FROM WANG DATASET

Metric	**Number of Génie neuroblastoma genes identified**	**Génie Score**	**Metric Rank**
Weighted Betweenness	21	6078	1st
Combined Metric	21	6077	2nd
Degree Centrality	21	5825	3rd
Weighted Degree	20	5600	4th
Eigenvector Centrality	20	5600	4th
Weighted Closeness	19	5587	6th

Applying the scoring to the top 500 genes identified by each metric, weighted betweenness is again the best performing metric, followed by the combined metric. Compared to the stage 4M network in the Molenaar dataset, the metrics identify fewer Génie neuroblastoma genes. Across the whole network, 89 out of the 500 neuroblastoma genes generated by Génie are present, compared to 150 in the stage 4M Molenaar network. Across all of the categories in both datasets, more Génie genes are present in three of the Molenaar category networks, and only one of the Wang category networks.

6.10 Metric Performance GRNs Inferred from Wang Dataset

Previously, each metric was ranked based on the top 20, top 100 and top 500 genes they identified from each neuroblastoma disease stage GRN for the Wang dataset. This ranking is based on the score of the identified genes in the Génie neuroblastoma gene list.

The results of these metric rankings can now be presented together, allowing an analysis of which metric is the best performing overall. Assigning an equal weighting to the scores of the top 20, top 100, and top 500 ranked genes for each metric in each category, the following ranks shown in table 6.53 below are given.

TABLE 6.53 OVERALL METRIC RANKS IN THE DIFFERENT NEUROBLASTOMA DISEASE STAGE GRNS INFERRED USING NOVEL INFERENCE METHOD FROM THE WANG DATASET

	Stage 1 Rank	**Stage 3 Rank**	**Stage 4 Rank**	**Stage 4M Rank**	**Ranking Totals**	**Overall Rank**
Degree Centrality	1st	1st	1st	3rd	6	1st
Weighted Degree	1st	2nd	3rd	4th	10	2nd
Combined Metric	1st	4th	4th	2nd	11	3rd
Weighted Betweenness	4th	6th	2nd	1st	13	4th
Eigenvector Centrality	6th	2nd	6th	4th	18	5th
Weighted Closeness	5th	5th	5th	6th	21	6th

As can be seen from the table above, degree centrality is the best performing metric overall, followed by weighted degree and the combined metric. It should be noted that degree centrality was the joint best performing metric for the Molenaar dataset, as was weighted degree centrality. There is a positive correlation of 0.61 between the overall metric ranks in both the datasets. If the rankings for just the top 20 genes in each category are taken into consideration, the following scores shown in table 6.54 are obtained.

TABLE 6.54 OVERALL METRIC RANKS IN THE DIFFERENT NEUROBLASTOMA DISEASE STAGE GRNS INFERRED USING NOVEL INFERENCE METHOD FROM THE WANG DATASET BASED ON TOP 20 GENES IDENTIFIED BY EACH METRIC

	Stage 1 Rank	**Stage 3 Rank**	**Stage 4 Rank**	**Stage 4M Rank**	**Ranking Totals**	**Overall Rank**
Degree Centrality	1st	1st	2nd	2nd	6	1st
Weighted Closeness	1st	4th	1st	2nd	8	2nd
Weighted Degree	5th	1st	2nd	2nd	10	3rd
Weighted Betweenness	1st	6th	2nd	1st	10	3rd
Combined Metric	1st	5th	2nd	2nd	10	3rd
Eigenvector Centrality	5th	3rd	6th	2nd	16	6th

Degree centrality is again the best performing metric, followed by weighted closeness centrality. There is a negative correlation of 0.4 with the metric ranks for the top 20 genes in the Molenaar dataset. If we take the ranking for the top 100 genes into account, the following results are obtained.

TABLE 6.55 OVERALL METRIC RANKS IN THE DIFFERENT NEUROBLASTOMA DISEASE STAGE GRNS INFERRED USING NOVEL INFERENCE METHOD FROM THE WANG DATASET BASED ON TOP 100 GENES IDENTIFIED BY EACH METRIC

	Stage 1 Rank	**Stage 3 Rank**	**Stage 4 Rank**	**Stage 4M Rank**	**Ranking Totals**	**Overall Rank**
Degree Centrality	2nd	2nd	1st	2nd	7	1st

Weighted Degree	1st	4th	2nd	2nd	9	2nd
Eigenvector Centrality	3rd	3rd	4th	2nd	12	3rd
Combined Metric	4th	1st	5th	2nd	12	3rd
Weighted Betweenness	6th	6th	3rd	1st	16	5th
Weighted Closeness	5th	5th	6th	2nd	18	6th

Once again, degree centrality is the best performing metric. Weighted degree centrality is the second best performing metric. These two metrics also rank in the exact same position, first and second respectively, for the top 100 in the Molenaar dataset, and there is a positive correlation of 0.63 between the metric ranks across the two datasets for the top 100. Finally, taking just the top 500 rankings into account, the following results shown in table 6.56 are obtained.

TABLE 6.56 OVERALL METRIC RANKS IN THE DIFFERENT NEUROBLASTOMA DISEASE STAGE GRNS INFERRED USING NOVEL INFERENCE METHOD FROM THE WANG DATASET BASED ON TOP 500 GENES IDENTIFIED BY EACH METRIC

	Stage 1 Rank	Stage 3 Rank	Stage 4 Rank	Stage 4M Rank	Ranking Totals	Overall Rank
Weighted Degree	1st	2nd	3rd	4th	10	1st
Degree Centrality	4th	2nd	2nd	3rd	11	2nd
Weighted Betweenness	3rd	6th	1st	1st	11	2nd
Combined Metric	2nd	4th	5th	2nd	13	4th

Eigenvector Centrality	6th	1st	4th	4th	15	5th
Weighted Closeness	5th	5th	6th	6th	22	6th

Weighted degree centrality just out-performs degree centrality, with both metrics showing similar performance. The eigenvector centrality is the best performing metric for the top 500 in the Molenaar dataset; here it is the 5th best performing metric. Between the top 500 in both datasets, there is a very weak positive correlation of 0.05.

6.11 Distribution of Genes in Combined Metric Top 20 Rankings List across the GRNs Inferred from Wang Datasets

As noted in the category sections, none of the genes that appear in the top 20 combined metric rankings list for any category appear in any other. This can be seen on the Venn diagram in figure B.10 of the appendix, showing the distribution of genes in the top 20 combined metric rankings list in the categories. Furthermore, extending this to look at the top 100 combined metric rankings lists, shown in the Venn diagram in figure B.11 of the appendix, this is also the case and suggests that this approach is suited to identifying specific genes associated with the different survival time

6.12 Overlap of Combined Metric Top 20 and Top 100 Ranked Genes in Common Categories

The final comparison between the two datasets is to see whether there is any overlap in the genes that are identified as being important in the neuroblastoma categories across the two datasets. It has been noted before that these two datasets are quite different, and as such, it might be expected that significant overlap does not occur. Two comparisons will be made;

firstly the top 20 combined metric ranking genes for the same category identified in the two datasets will be plotted on a Venn diagram, figures B.12 – B.15 in the appendix, and secondly the top 100 combined metric ranking genes will be plotted against each other, figures B.16 – B.19 in the appendix.

As can be seen from figures B.12 and B.13, there are no genes in common for either stage 1, or stage 3, across both datasets. There is one gene common to the stage 4 categories in both categories, CDH4. Whilst not specifically identified in previous neuroblastoma studies, in a 2003 study by Charrasse et al [136] it was shown to be up-regulated in the paediatric solid tumour Rhabdomyosarcoma.

There is quite a significant overlap in the stage 4M categories across both datasets, shown in figure B.15. 7 out of the top 20 ranked genes for the stage 4M category network in each dataset are also ranked in the top 20 in the other dataset. Amongst these seven genes in common, the presence of MYCN should be highlighted. The high number of genes in common across the same neuroblastoma category is perhaps also a surprising finding, considering the differences in the datasets already noted.

Extending the number of genes to the top 100 ranked genes in each stage, there are more genes in common across the stages in the two datasets. There are 4 genes common to stage 1; TGDS, EPN2, EPB41L3, and DLG4. None of these genes appear on the top 500 genes returned by Génie for neuroblastoma. There is only one gene common to stage 3 across the Wang and Molenaar datasets, GPC5, which has been the subject of lung cancer studies, but none relating specifically to neuroblastoma.

There are eight genes in common to both stage 4 categories; PDCD6, GGA3, CHGA, CDH4, BZRAP1, ARFIP2, APOBEC3B, and AP1G2. None of these eight genes appear on the top

500 genes returned by Génie for neuroblastoma. Finally, there are 23 genes common to both stage 4M categories; TRAP1, TKT, TGIF2, SNRPA, RUVBL1, RSL1D1, PFAS, PES1, PAICS, NMU, MYCN, LAPTM4B, GPR125, GMPS, FBL, DDX10, DDX1, CBS, CAD, BYSL, BAZ1A, BAMBI, and APEX1. Of these genes, only the number 1 ranked MYCN appears on the top 500 genes returned by Génie for neuroblastoma.

6.13 Discussion

Having used the Z score method to infer networks for different categories in the two datasets, a number of concluding observations can be made. Firstly, as suggested by the initial correlation of the gene expression level rankings, the datasets differ substantially. This is highlighted in both the genes that are highly ranked for each neuroblastoma stage in the datasets, and also in the genes that comprise the largest unique genes in the cliques. It is important to take the different array platforms used in the different studies into account; the Wang study used the relatively old Affymetrix U95 microarray design with approximately 10,000 genes represented, whilst the Molenaar study used the far more recent Affymetrix EXON ST 1.0 design with approximately 20,000 genes represented. There are also a number of technical differences in relation to how the arrays are constructed, and how background is calculated.

There are also differences in the network level metrics, with only the diameter values correlating exactly. However, there were a number of common genes identified in the stage 4M categories for both datasets. This is a promising finding, although perhaps it should be remembered that this was the only category specifically classed based on a genomic property, high MYCN amplification, which might go some way to explaining why this is the case.

Looking at the cliques, there are a number of enrichment results of note for the genes that make up the largest unique cliques in the network inferred from the Molenaar dataset. In particular, the 10th enrichment result for both stage 2 and stage 4M are of particular interest, relating to specific biological processes previously observed in neuroblastoma, and also corresponding to the neuroblastoma stage. Three out of the four largest unique cliques inferred for the Molenaar dataset had results relating to neuroblastoma in their respective top ten enrichment results, suggestive of biological relevance and significance.

For the list of genes that make up the largest unique cliques in the networks inferred from the Wang dataset, there is only one result of note, detailing SP1 gene regulation. The list of genes that comprise the largest unique clique in the stage 4 network did not return any enrichment results, and the list of genes that comprise the largest unique clique in the stage 1 network returned one enrichment result. This would suggest that the cliques in the networks in the Wang dataset are not involved in significant biological processes, and in particular, are not involved in biological processes associated with neuroblastoma. Comparing the cliques in the networks inferred from the two datasets, it would appear that the cliques in the networks inferred from the Molenaar dataset are biologically of greater interest. This is based on the greater number of enrichment results returned specifically relating to neuroblastoma.

Of particular note is the top ranking of the gene MYCN in both the GRN inferred from the stage 4M samples in the Wang dataset, and also the Molenaar dataset. The feedback from biologists relating to the top ranked genes in the glioblastoma survival categories has been previously noted. Prior to constructing the GRNs for the neuroblastoma disease stages, one biologist specifically indicated that she would expect to see the gene MYCN as the top rated gene for the 4M stage networks. The fact that this is the case for the networks in both of the datasets not only suggests that the network inference method is capable of inferring networks that are biologically meaningful, but also are capable of inferring networks that biologists have confidence in. This was not the case with the networks inferred from the glioblastoma dataset using the WGCNA method. It is important to note however that the distribution of MYCN expression values is somewhat different from the majority of other genes. It tends to be either very high due to gene amplification, or quite low. In contrast, most other genes are either consistently high or consistently low in all samples, or have a much broader distribution from low to high. As a result, MYCN will have a much greater discriminatory power than most other genes; particularly since it is also strongly associated with a specific clinical sub-type implying biological significance, in this case neuroblastoma stage 4M.

Overall, the results from this chapter give an insight into some of the issues that exist in replicating results across different datasets. There have been a number of proposed approaches to deal with this problem, such as different normalisation techniques, however this is still an area without an adequate solution. Another problem is the use of non-genomic criteria for categorising microarray data. Despite all this, there are a number of promising results. As well as MYCN being the top ranked genes in both GRNs inferred from the stage 4M samples in both datasets, there are a number of common genes in the top ranking genes in these two networks. If the Spearman rank score of 0.63 between the ranking of the common gene average expression values is taken into account as well between the stage 4M samples in both datasets, this illustrates that this method has been able to identify common top ranking genes in the stage 4M categories across both datasets despite the obvious differences in the gene expression values. Whilst on the whole this chapter has shown the problems that exist replicating results across microarray datasets, there is a significant overlap in the stage 4M results, and demonstrates that using a genomic criteria for classifying data can result in some degree of replication across different datasets.

6.14 Conclusions

To conclude, in this chapter the novel inference method introduced in the last chapter was applied to two different neuroblastoma microarray datasets from two different studies. All three objectives were addressed in this chapter. Firstly, the transferability of the novel inference method proposed in the previous chapter was demonstrated through applying it to two neuroblastoma microarray datasets. Secondly, various metrics were used to identify different genes in the different disease stage neuroblastoma GRNs across both datasets, and the performance of the metrics was compared and analysed. Thirdly, the different disease stage neuroblastoma GRNs across both datasets were compared, and genes common to the same disease stage GRN across both neuroblastoma microarray datasets were identified. In the next chapter, a further refinement of the novel network inference method is presented, and it is applied to a retinoblastoma microarray dataset. In the next chapter, a further refinement of the novel inference method is applied to a proprietary retinoblastoma microarray dataset. Samples from normal retinal data are contained in this dataset, and as such, a sample reference network for normal retinal data is also constructed to compare to the retinoblastoma networks.

Chapter 7

Application of Refined Novel Network Inference Method to retinoblastoma dataset

In this chapter, a further refined version of the novel network inference method used in chapters 5 and 6 is presented. This is done with the purpose of implementing a GRN inference method that is able to distinguish between genes that are amplified or under-expressed together, following feedback from biologists. All of the datasets analysed so far in the work have been taken from previously published studies containing samples from different stages of disease that do not contain healthy reference data for comparison. In this chapter, the analysis of a proprietary retinoblastoma microarray dataset is carried out that also has normal samples allowing a reference normal GRN to be inferred. This dataset is from a very recent 2013 study by Kapatai et al [137].

7.1 Retinoblastoma Microarray Dataset

Retinoblastoma is an aggressive cancer of the retina that arises due to a mutation on the RB1 gene at chromosome 13 [138, 139]. The retinoblastoma dataset consists of 21 samples from Birmingham Children's Hospital, BCH, analysed using the Affymetrix HuGene Array platform. The samples have been classified into three groups by researchers at BCH; blue, red, and green. Two of the samples were found to be similar to normal retina, and subsequently lead to the assumption that normal rather than tumour tissue was supplied. As such, these two samples are taken as normal samples from which the normal reference GRN will be inferred. For the rest of the samples, standard analysis methods including SAM and

PCA were used to detect two different retinoblastoma groups, red, and blue. There were some clinical associations relating to tumour aggressiveness that suggested the blue category of samples is a more advanced and aggressive retinoblastoma category than the red category. It should be noted that both these groups contain high stage retinoblastoma samples, as the only retinoblastoma tissue available for investigation is from high stage tumours due to these being the the only ones which are removed surgically. There are 13 samples in the blue category, six in the red category, and two in the green category. For the rest of the chapter, the categories will be referred to as RB Blue for the category containing the blue samples, RB Red for the category containing the red samples, and RB Green for the category containing the green samples.

As before, prior to inferring networks from the different categories in the microarray dataset, an idea of how similar the categories are can be gauged from how well the gene expression levels correlate. Table 7.1 below shows the correlation scores between the rankings of the gene expression levels.

TABLE 7.1 CORRELATION SCORES OF AVERAGE GENE EXPRESSION RANKINGS IN RETINOBLASTOMA MICROARRAY DATASET

Category	RB Blue	RB Red	RB Green
RB Blue	1	0.9277262	0.8775253
RB Red	0.9277262	1	0.9285888
RB Green	0.8775253	0.9285888	1

From the table above, it can be seen that the greatest discrepancy between the gene expression rankings occurs between the RB Blue and RB Green categories. This corresponds to the biology; RB Blue is the most advanced retinoblastoma category, and RB Green containing samples from normal retinal data. It would therefore be expected that the gene expression levels between these two categories would be those with the greatest difference.

Instead of the absolute value Z score approach used in the two previous chapters, a further modified network inference technique will be used in this chapter. It is beneficial to be able to further understand the relationship between gene expression levels by identifying which genes are amplified together, and which genes are under-expressed together. In order to ascertain this, two additional approaches will be implemented. A positive-positive variant of the Z score approach, where only positive Z scores greater than 0 are retained, and a negative-negative variant of the Z score approach where only negative Z scores less than 0 are retained. The positive-positive variant will be referred to as the positive Z score method, and the negative-negative variant will be referred to as the negative Z score method.

7.2 Network Level Metrics in the Retinoblastoma GRNs Inferred using the Positive Z Score Method

As was the case in the last chapter, prior to identifying individual genes in the networks using node level metrics, the network level metrics will be calculated and analysed. The scores for these are shown in table 7.2. The same metrics will be used again; weighted degree assortativity, degree asssortativity, diameter, and network clustering. The largest unique cliques will again be calculated for each of the networks for the two different datasets, and the genes that comprise these cliques will be investigated for both enrichment and the presence of retinoblastoma genes of interest using Génie. As was the case with neuroblastoma, the list generated by Génie is extensive, and only the top 500 ranked genes for retinoblastoma will be used. The Génie list of scoring genes for retinoblastoma is shown in table A.35 of the appendix.

TABLE 7.2 NETWORK LEVEL SCORES ACROSS THE CATEGORIES FOR THE RETINOBLASTOMA GRNS INFERRED USING THE POSITIVE Z SCORE METHOD

Metric	**RB Blue**	**RB Red**	**RB Green**
Weighted Degree Assortativity	0.068	-0.268	-0.257
Degree Assortativity	0.103	-0.262	-0.248
Diameter	4.352	3.959	3.263
Network Clustering	0.652	0.437	0.844
Largest Clique Size	132	96	238

Looking at the network level scores, the value of the diameter is of interest. In the previous chapter, it was noted that in both the neuroblastoma datasets that the value of the diameter was greatest in the most advanced neuroblastoma category. Here, in the most advanced retinoblastoma category, the diameter value is also greatest. This could again be an indication that the diameter is a network level property that is associated with disease stage, and more specifically, one could hypothesise that this could be due to genes involved in certain cellular processes not being able to easily communicate with each other, thereby contributing to the more advanced stage of disease. The size and composition of the largest unique cliques in the different categories can be informative; this is looked at in the next section.

7.3 Largest Clique Calculation and Analysis in Retinoblastoma GRNs Inferred using Positive Z Score Method

As well as identifying common and unique genes in the cliques for the different categories, it is informative to see whether these genes have been previously identified as being retinoblastoma genes of interest and appear on the list of Génie retinoblastoma genes. The list of genes that comprise the largest unique clique in each network category can also be checked for enrichment using the GenesetDB website. Table 7.3 below shows the size of the cliques in each category, whether any of the genes that comprise this clique are present in any other cliques, and finally whether any of the genes in the clique appear on the Génie retinoblastoma gene list.

TABLE 7.3 RETINOBLASTOMA GENES OF INTEREST IN THE LARGEST UNIQUE CLIQUES IN THE RETINOBLASTOMA GRNS INFERRED USING POSITIVE Z SCORE METHOD

	Clique Size	Genes present in other RB cliques	Number of Génie genes and score
RB Blue	132	0	7, score of 1909
RB Red	96	0	3, score of 1323
RB Green	238	0	11, score of 3207

Across the cliques, a number of retinoblastoma genes that appear on the list generated by Génie are present. The RB Green category has both the highest number of Génie genes, and also the highest score of genes identified. This might appear to be a surprising result, considering that the RB Green category network has been inferred from healthy retinal samples. One possible explanation for this is that genes identified by Génie for a disease might also be responsible for regulation of healthy cellular processes. Also, unfortunately Génie does not allow a distinction to be made as to whether genes that are present of the list have been identified in the studies as being significantly over or under-expressed, another potential explanation. The RB Red category has the lowest number of genes identified, and the lowest score. Only 21 genes that appear on the list generated by Génie for retinoblastoma are present in the cliques, which represents just 4.2 % of the complete list used. This lower proportion though can be attributed to the more specific network inference method used; it will be of interest to compare the cliques that are identified in the two different network inference techniques to see the composition of the cliques for the negative Z score method. The Venn diagram in figure B.20 of the appendix shows the distribution of the genes in the largest unique cliques in the networks inferred using the positive Z score method.

There are 466 genes in total across the three cliques. There are no genes common to either two or three cliques, with each gene only appearing in one clique. This result is indicative that the cellular processes are very different in the cliques in the different retinoblastoma categories. It would be expected that different genes are over-expressed in different retinoblastoma stages and healthy cellular processes; the distribution of genes in the cliques corresponds with this.

Having investigated the distribution of genes in the cliques, the next step is to investigate the enrichment of these genes using GenesetDB. Once again, these tables will be presented in the appendix. For the genes that comprise the largest unique clique in the RB Blue category, only seven results are returned from GenesetDB. This would imply that this combination of genes has not been identified as being involved in biological processes. The seven results returned are shown in table A.36 of the appendix.

The genes that comprise the largest unique clique in the RB Red category do return any enrichment results. As was the case with the RB Blue category, this would imply that this combination of genes has not been identified as being involved in biological processes.

Unlike the previous two categories, the genes that comprise the largest unique clique in the RB Green category return a number of enrichment results. The top ten results are shown in table A.37 of the appendix. As noted before, it might be surprising that the network inferred from healthy retinal data returns the most biologically relevant results. However, it should be remembered that normal tissues are highly specialized to carry out specific functions, and as a result have a gene expression profile which reflects this. This can be seen in the results; there are a number of results that relate to eye function and degradation. As noted with Génie, the enrichment results to do not specify whether these results are associated with increased or decreased gene expression. This accounts for the presence of the first result, visual perception, and also the 5th result, retinal degeneration. As might be expected, genes involved in visual perception are highly expressed in normal retinal tissue [140]; therefore in a GRN inferred from normal tissue using the positive Z score method, enrichment results relating to visual perception are present. These results are also a reminder of the nature of the work in the thesis; it is intended to be able to guide clinicians and biologists with specialist and specific medical and biological knowledge that can use this knowledge to interpret the results.

7.4 RB Blue Category GRN Inferred using Positive Z Score Method

Having investigated the network level metrics, and the largest unique cliques in the networks, the focus now is using node level metrics to identify genes of interest. As before, a combined metric comprised of weighted degree, weighted betweenness and weighted closeness, is used to infer a list of top 20 ranked genes for this disease category, and all subsequent categories.

Enrichment and the presence of previously identified retinoblastoma genes will be investigated for the combined list. Correlations will be calculated for the metrics, and the metrics will be scored using the 500 top ranked genes for retinoblastoma returned by Génie.

Table 7.4 shows the top 20 ranked genes for the combined metric in the RB Blue category GRN inferred using the positive Z score method. Four of these genes in the table appear in the list of 500 retinoblastoma genes generated by Génie; EYA1, MYC, HMGA2, and CD24. This represents quite a high proportion of genes having been previously identified as interest for retinoblastoma. None of the genes that appear in the top 20 ranked list appear in any of the top 20 ranked for the other categories. As was the case with the largest unique cliques, this suggests clearly distinguishable differences between the retinoblastoma categories in terms of the genes that are playing important roles in cellular processes.

Despite 4 genes out of the 20 appearing on the Génie retinoblastoma list, no enrichment results are returned for the 20 genes in the table below. This suggests that despite a number of these genes being associated with retinoblastoma, they have not been identified as being involved together in biological processes.

TABLE 7.4 TOP 20 RANKED GENES FOR COMBINED METRIC IN RB BLUE GRN INFERRED USING POSITIVE Z SCORE METHOD

Gene	*Combined Metric Rank*	*Gene*	*Combined Metric Rank*
EYA1	1	HMGA2	11
MYC	2	IGHD	12
DSC2	3	TRPC5	13
DPP10	4	SOX4	14

KCNQ5	5	RSPO1	15
ZNF193	6	NEUROG1	16
IGLJ3	7	OR4C16	17
RPS24	8	MED20	18
RPS27A	9	TFF1	19
TUBB2B	10	CD24	20

Table 7.5 below shows the Spearman Rank correlation scores for the rankings assigned to the genes by each metric in the RB Blue GRN inferred using the positive Z score method.

TABLE 7.5 SPEARMAN RANK CORRELATION SCORES OF RANKINGS ASSIGNED BY THE METRICS TO THE GENES IN RB BLUE GRN INFERRED USING POSITIVE Z SCORE METHOD

	Degree Centrality	**Weighted Degree**	**Eigenvector Centrality**	**Closeness Centrality**	**Betweenness Centrality**
Degree Centrality	1	0.999	0.415	0.687	0.572
Weighted Degree	0.999	1	0.430	0.699	0.575
Eigenvector Centrality	0.415	0.430	1	0.562	0.176
Closeness Centrality	0.687	0.699	0.562	1	0.732
Betweenness Centrality	0.572	0.575	0.176	0.732	1

From table 7.5 above, it can be seen that for this category, and unlike most of the previous categories in this work to date, there are a number of weak correlations. In particular, the weak correlation of 0.177 between the eigenvector centrality and the weighted betweenness centrality stands out. There are also a number of other correlations with scores less than 0.5. These weak correlations imply that there is not as substantial an overlap in the genes that the metrics rank highly as was the case with the other categories, and that the genes that the metrics identify will be quite different.

The list of genes generated by Génie for retinoblastoma will be used to score the metrics, limited to the top 500 ranking genes. This is due to a great deal of genes returned, and some specificity of results is desired. Starting with the top 20 ranked genes by each metric, table 7.6 below shows the scores that are obtained.

TABLE 7.6 GÉNIE SCORES AND METRIC RANKING FOR THE TOP 20 RANKED GENES FOR EACH METRIC IN RB BLUE GRN INFERRED USING POSITIVE Z SCORE METHOD

Metric	**Number of Génie retinoblastoma genes identified**	**Génie Score**	**Metric Rank**
Weighted Betweenness	4	1581	1st
Combined Metric	4	1403	2nd
Weighted Closeness	2	646	3rd
Degree Centrality	1	150	4th
Weighted Degree	1	150	4th
Eigenvector Centrality	1	150	4th

The weighted betweenness centrality is the best performing metric, followed by the combined metric. Both the weighted betweenness centrality and the combined metric identify four genes, although the weighted betweenness identifies higher scoring genes. Applying the scoring system to the top 100 ranked genes for each metric, the results in table 7.7 below are given.

TABLE 7.7 GÉNIE SCORES AND METRIC RANKING FOR THE TOP 100 RANKED GENES FOR EACH METRIC IN RB BLUE GRN INFERRED USING POSITIVE Z SCORE METHOD

Metric	**Number of Génie retinoblastoma genes identified**	**Génie Score**	**Metric Rank**
Combined Metric	7	2546	1st
Weighted Closeness	7	2327	2nd
Weighted Betweenness	5	1631	3rd
Weighted Degree	5	1592	4th
Degree Centrality	4	1122	5th
Eigenvector Centrality	3	946	6th

Applying the scoring system to the top 100 genes identified by the metrics, the combined metric is the best performing metrics, followed by the weighted closeness centrality. Both these metrics identify 7 genes, with the genes identified by the combined metric being higher scoring. The eigenvector centrality metrics perform relatively poorly, only identifying 3 genes. Finally, extending the scoring system to the top 500 genes, the following results shown in table 7.8 below are given.

TABLE 7.8 GÉNIE SCORES AND METRIC RANKING FOR THE TOP 500 RANKED GENES FOR EACH METRIC IN RB BLUE GRN INFERRED USING POSITIVE Z SCORE METHOD

Metric	**Number of Génie retinoblastoma genes identified**	**Génie Score**	**Metric Rank**
Eigenvector Centrality	25	7062	1st
Degree Centrality	17	5192	2nd
Weighted Degree	17	5192	2nd

Weighted Closeness	17	4846	4th
Combined Metric	14	4078	5th
Weighted Betweenness	12	3408	6th

Despite performing badly in the two previous categories, this time the eigenvector centrality performs best, clearly outperforming the other metrics. The weighted degree and degree centrality are the next best performing metrics. Three of the top 25 ranked genes by Génie for retinoblastoma are identified by these metrics; TP53 ranked 3rd, MYCN ranked 6th, and KIT ranked 25th. This implies that these metrics are able to identify biologically relevant genes of interest in the network inferred from the RB Blue samples in the retinoblastoma dataset.

Across the whole network, 48 out of the 500 retinoblastoma genes generated by Génie are present, with 25 of these present in the top 500 based on eigenvector centrality. However, only 4 out of the top 50 Génie ranked retinoblastoma genes are present in the top 500 ranked retinoblastoma genes for any of the metrics. This shows that many of the genes that Génie highlights as being important for retinoblastoma are not identified by the network metrics in this category. This could be due to three reasons. Firstly, that the network inference method and metrics are not accurate in re-constructing and identifying the gene regulatory processes that take place; secondly, that the modified version of the Z score method that has been applied only identifies a small number of retinoblastoma genes of interest; thirdly, that the genes that are present on the list generated by Génie do not in fact play an important role in this retinoblastoma category. It should again be noted that as well as no information concerning under and over-expressed genes, there is also no stage specific information present with the genes identified by Génie, i.e. specific genes are not associated with specific retinoblastoma stages. One possible reason for this is that the only retinoblastoma tissue available for investigation is from high stage tumours, as these are the only ones which are removed surgically. As a result, there is very little genetic information about low stage tumours, except in the very few cases where these are detected in eyes simultaneously with a higher stage tumour.

7.5 RB Red Category GRN Inferred using Positive Z Score Method

Table 7.9 shows the top 20 ranked genes for the combined metric in the RB Red category GRN inferred using the positive Z score method. None of the genes in the table appear in the list of 500 retinoblastoma genes generated by Génie. Also, none of the genes that appear in the top 20 ranked list for the combined metric appear in any of the top 20 combined metric ranked lists for the other categories. As was the case with the previous category, this suggests that different genes are playing important roles in the different retinoblastoma stages.

In contrast to the last category, where a number of the top 20 combined metric ranked genes appeared on the Génie list but no enrichment results were returned, a number of enrichment results are returned for this list of genes. The top ten results are shown in table A.38 of the appendix. However, none of these enrichment results directly relate to retinal biology or retinoblastoma.

TABLE 7.9 TOP 20 RANKED GENES FOR COMBINED METRIC IN RB RED GRN INFERRED USING POSITIVE Z SCORE METHOD

Gene	***Combined Metric Rank***	***Gene***	***Combined Metric Rank***
PGM3	1	SNUPN	11
ACAT1	2	EXOSC8	12
C20ORF72	3	AKR1A1	13

TRPM7	4	GPRASP2	14
SESN1	5	TBC1D5	15
YARS	6	TRIM32	16
C6ORF204	7	NEDD4	17
CARS2	8	TEX9	18
ZNF702P	9	PDHB	19
FBLN5	10	SGPL1	20

Table 7.10 below shows the Spearman Rank correlation scores for the rankings assigned to the genes by each metric in the RB Red GRN inferred using the positive Z score method.

TABLE 7.10 SPEARMAN RANK CORRELATION SCORES OF RANKINGS ASSIGNED BY THE METRICS TO THE GENES IN RB RED GRN INFERRED USING POSITIVE Z SCORE METHOD

	Degree Centrality	**Weighted Degree**	**Eigenvector Centrality**	**Closeness Centrality**	**Betweenness Centrality**
Degree Centrality	1	0.999	0.953	0.903	0.899
Weighted Degree	0.999	1	0.949	0.897	0.900
Eigenvector Centrality	0.953	0.949	1	0.980	0.799
Closeness Centrality	0.903	0.897	0.980	1	0.751
Betweenness Centrality	0.899	0.900	0.799	0.751	1

Unlike the previous category, where there were a number of weak correlations, in the table above it can be seen that for this category all of metrics have a positive correlation greater than 0.75. These strong correlations imply that there is a substantial overlap in the genes that the metrics rank highly and we would expect that there will be quite a lot of overlap in the genes that these metrics identify.

Applying the scoring system to the top 20, top 100, and top 500 genes identified by each of metrics, the metrics can be ranked. The scores for the top 20 genes identified by each metric are shown in table 7.11 below.

TABLE 7.11 GÉNIE SCORES AND METRIC RANKING FOR THE TOP 20 RANKED GENES FOR EACH METRIC IN RB RED GRN INFERRED USING POSITIVE Z SCORE METHOD

Metric	**Number of Génie retinoblastoma genes identified**	**Génie Score**	**Metric Rank**
Weighted Betweenness	1	454	1st
Degree Centrality	0	0	2nd
Weighted Degree	0	0	2nd
Weighted Closeness	0	0	2nd
Eigenvector Centrality	0	0	2nd
Combined Metric	0	0	2nd

As noted previously, no genes on the Génie top 500 ranked list are identified by the combined rank, and this is also the case for all of the metrics apart from the weighted betweenness centrality, which only identifies one gene despite being the best performing metric. Applying the scoring system to the top 100 ranked genes for each metric, the results in table 7.12 below are given.

TABLE 7.12 GÉNIE SCORES AND METRIC RANKING FOR THE TOP 100 RANKED GENES FOR EACH METRIC IN RB RED GRN INFERRED USING POSITIVE Z SCORE METHOD

Metric	**Number of Génie retinoblastoma genes identified**	**Génie Score**	**Metric Rank**
Eigenvector Centrality	3	961	1st
Degree Centrality	3	925	2nd
Weighted Degree	3	925	2nd
Weighted Closeness	3	925	2nd
Combined Metric	3	925	2nd
Weighted Betweenness	2	572	6th

Applying the scoring system to the top 100 genes identified by the metrics, the eigenvector centrality is the best performing metric, identifying three genes. All of the other metrics apart from the weighted betweenness also identify three genes, but these are lower scoring. One of the three genes identified by the eigenvector centrality, RBL2, is the 11th ranked gene for retinoblastoma. Fewer genes are identified by the metrics than for the last retinoblastoma stage. Finally, extending the scoring system to the top 500 genes, the following results shown in table 7.13 below are given.

TABLE 7.13 GÉNIE SCORES AND METRIC RANKING FOR THE TOP 500 RANKED GENES FOR EACH METRIC IN RB RED GRN INFERRED USING POSITIVE Z SCORE METHOD

Metric	**Number of Génie retinoblastoma genes identified**	**Génie Score**	**Metric Rank**
Degree Centrality	18	5481	1st
Weighted Degree	18	5481	1st

Combined Metric	18	5481	1st
Eigenvector Centrality	18	5465	4th
Weighted Betweenness	17	5016	5th
Weighted Closeness	17	5003	6th

The combined metric, weighted degree and degree centrality are the joint best performing metrics, identifying the same 18 genes. Eigenvector centrality also identifies 18 genes, but these are lower scoring. Five of the top 25 ranked genes for retinoblastoma are identified by the joint best performing metrics, compared to 3 of the top 25 by the best performing metric for the previous retinoblastoma stage. Extending this, 6 of the top 50 ranked genes are identified, again better than the best performing metric for the last category. Across the whole network, 53 out of the 500 retinoblastoma genes generated by Génie are present, slightly more than were identified in the last category.

7.6 RB Green Category GRN Inferred using Positive Z Score Method

Table 7.14 shows the top 20 ranked genes for the combined metric in the RB Green category GRN inferred using the positive Z score method. As was the case with the previous category, none of the genes in the table appear in the list of 500 retinoblastoma genes generated by Génie. Also, none of the genes that appear in the top 20 combined metric ranked list appear in any of the top 20 combined metric ranked lists for the other categories. This again suggests that different genes are playing important roles in the different retinoblastoma stages.

Also in keeping with the last category, a number of enrichment results are returned for this list of genes. The top ten results are shown in table A.39 of the appendix. However, unlike the last category, amongst these enrichment results there is one biological process that involves the retina. Rods, named in the first enrichment result, are photoreceptor cells in the retina.

TABLE 7.14 TOP 20 RANKED GENES FOR COMBINED METRIC IN RB GREEN GRN INFERRED USING POSITIVE Z SCORE METHOD

Gene	***Combined Metric Rank***	***Gene***	***Combined Metric Rank***
GNAT1	1	CLUL1	11
RHO	2	ANK3	12
PDE6G	3	PDE8B	13
CALB2	4	UNC13C	14
NFIA	5	PVALB	15
WIF1	6	MAPK10	16
LSAMP	7	SNORD115-32	17
GPR37	8	C1ORF61	18
GABRA1	9	TANC1	19
SEMA3A	10	NT5E	20

Table 7.15 below shows the Spearman Rank correlation scores for the rankings assigned to the genes by each metric in the RB Green GRN inferred using the positive Z score method.

TABLE 7.15 SPEARMAN RANK CORRELATION SCORES OF RANKINGS ASSIGNED BY THE METRICS TO THE GENES IN RB GREEN GRN INFERRED USING POSITIVE Z SCORE METHOD

	Degree Centrality	**Weighted Degree**	**Eigenvector Centrality**	**Closeness Centrality**	**Betweenness Centrality**
Degree Centrality	1	0.998	0.993	0.991	0.428
Weighted Degree	0.998	1	0.996	0.984	0.424
Eigenvector Centrality	0.993	0.996	1	0.974	0.394
Closeness Centrality	0.991	0.984	0.974	1	0.446
Betweenness Centrality	0.428	0.424	0.394	0.446	1

The correlation scores for this category make for interesting reading. It is a mix of extremely strong correlations between almost all of the metrics apart from the weighted betweenness centrality, and weak correlations between the weighted betweenness centrality and the other metrics. As such, four of the five metrics would be expected to identify almost identical genes as being highly ranked. The genes identified by the weighted betweenness would be expected to be somewhat different though.

Applying the scoring system to the top 20, top 100, and top 500 genes identified by each of metrics, the metrics can be ranked. The scores for the top 20 genes identified by each metric are shown in table 7.16 below.

TABLE 7.16 GÉNIE SCORES AND METRIC RANKING FOR THE TOP 20 RANKED GENES FOR EACH METRIC IN RB GREEN GRN INFERRED USING POSITIVE Z SCORE METHOD

Metric	**Number of Génie retinoblastoma genes identified**	**Génie Score**	**Metric Rank**
Weighted Degree	2	680	1st
Eigenvector Centrality	2	680	1st
Degree Centrality	1	432	3rd
Weighted Closeness	1	432	3rd
Weighted Betweenness	1	125	5th
Combined Metric	0	0	6th

The weighted degree and eigenvector centrality are the best performing metrics, identifying the same two genes, followed by the degree and weighted closeness that both identify the same gene. Applying the scoring system to the top 100 ranked genes for each metric, the results in table 7.17 below are given.

TABLE 7.17 GÉNIE SCORES AND METRIC RANKING FOR THE TOP 100 RANKED GENES FOR EACH METRIC IN RB GREEN GRN INFERRED USING POSITIVE Z SCORE METHOD

Metric	**Number of Génie retinoblastoma genes identified**	**Génie Score**	**Metric Rank**
Weighted Closeness	6	1841	1st
Combined Metric	6	1841	1st
Weighted Degree	6	1794	3rd
Eigenvector Centrality	6	1794	3rd
Degree Centrality	5	1537	5th

Weighted Betweenness	5	1073	6th

The weighted closeness and combined metric are the joint best performing, identifying the same six genes. The weighted degree and eigenvector centrality also identify six genes, but these are lower scoring. Finally, extending the scoring system to the top 500 genes, the following results shown table 7.18 below are given.

TABLE 7.18 GÉNIE SCORES AND METRIC RANKING FOR THE TOP 500 RANKED GENES FOR EACH METRIC IN RB GREEN GRN INFERRED USING POSITIVE Z SCORE METHOD

Metric	**Number of Génie retinoblastoma genes identified**	**Génie Score**	**Metric Rank**
Degree Centrality	27	7541	1st
Weighted Degree	27	7541	1st
Weighted Betweenness	27	7541	1st
Weighted Closeness	27	7541	1st
Eigenvector Centrality	27	7541	1st
Combined Metric	27	7541	1st

The metrics perform better at identifying genes in the Génie list of ranked retinoblastoma genes from their respective top 500 ranked genes for this category than for the two previous retinoblastoma categories. This is borne out by all 27 of the Génie ranked genes present in the network being identified. Previously, it was suggested that due to the extremely strong correlation scores between some of the metrics, we would expect to see the metrics identifying similar genes. This is the case here, although it was not expected that the weighted betweenness would also identify exactly the same genes as the other metrics. The lower number of retinoblastoma genes on the Génie list in the network is perhaps also expected, considering that this network has been inferred from what have been taken to be samples from normal retinal tissue.

7.7 Metric Performance in the GRNs Inferred from the Retinoblastoma Microarray using the Positive Z Score Method

Previously, each metric was ranked based on the top 20, top 100 and top 500 genes they identified from each retinoblastoma category GRN inferred using the positive Z score method. This ranking is based on the score of the identified genes in the Génie retinoblastoma gene list.

The results of these metric rankings can now be presented together, allowing an analysis of which metric is the best performing overall. Assigning an equal weighting to the scores of the top 20, top 100, and top 500 ranked genes for each metric in each category, the following ranks shown in table 7.19 below are given.

TABLE 7.19 OVERALL METRIC RANKS IN THE DIFFERENT RETINOBLASTOMA CATEGORY GRNS INFERRED FROM RETINOBLASTOMA MICROARRAY USING POSITIVE Z SCORE APPROACH

	RB Blue Category Rank	RB Red Category Rank	RB Green Category Rank	Ranking Totals	Overall Rank
Weighted Degree	3rd	1st	1st	5	1st
Combined Metric	1st	1st	4th	6	2nd
Weighted Closeness	2nd	5th	1st	8	3rd
Eigenvector Centrality	5th	4th	1st	10	4th

Degree Centrality	5th	1st	5th	11	5th
Weighted Betweenness	3rd	6th	6th	15	6th

Weighted degree is the best performing metric overall, followed by the combined metric and the weighted closeness centrality. As pointed out previously, the number of connections a node has in a network is the most often applied property for determining its importance in a network. It should perhaps not be surprising then to see a degree based metric perform well, although the poor performance of the degree centrality itself compared to the weighted degree perhaps is. If the rankings for just the top 20 genes in each category are taken into consideration, the following scores shown in table 7.20 are obtained.

TABLE 7.20 OVERALL METRIC RANKS IN THE DIFFERENT RETINOBLASTOMA CATEGORY GRNS INFERRED FROM RETINOBLASTOMA MICROARRAY USING POSITIVE Z SCORE APPROACH BASED ON THE TOP 20 GENES IDENTIFIED BY EACH METRIC

	RB Blue Category Rank	RB Red Category Rank	RB Green Category Rank	Ranking Totals	Overall Rank
Weighted Degree	4th	2nd	1st	7	1st
Weighted Betweenness	1st	1st	5th	7	1st
Eigenvector Centrality	4th	2nd	1st	7	1st
Weighted Closeness	3rd	2nd	3rd	8	4th
Degree Centrality	4th	2nd	3rd	9	5th
Combined Metric	2nd	2nd	6th	10	6th

Weighted betweenness, eigenvector centrality, and the weighted degree are the joint best performing metrics. The degree centrality again performs poorly. Taking the rankings for the top 100 genes into account, the following results shown in table 7.21 are obtained.

TABLE 7.21 OVERALL METRIC RANKS IN THE DIFFERENT RETINOBLASTOMA CATEGORY GRNS INFERRED FROM RETINOBLASTOMA MICROARRAY USING POSITIVE Z SCORE APPROACH BASED ON THE TOP 100 GENES IDENTIFIED BY EACH METRIC

	RB Blue Category Rank	**RB Red Category Rank**	**RB Green Category Rank**	**Ranking Totals**	**Overall Rank**
Combined Metric	1st	2nd	1st	4	1st
Weighted Closeness	2nd	2nd	1st	5	2nd
Weighted Degree	4th	2nd	3rd	9	3rd
Eigenvector Centrality	6th	1st	3rd	10	4th
Degree Centrality	5th	2nd	5th	12	5th
Weighted Betweenness	3rd	6th	6th	15	6th

This time, the combined metric is the best performing metric, followed by weighted closeness centrality and weighted degree. Finally, taking the top 500 gene into account, the following results in table 7.22 are obtained.

TABLE 7.22 OVERALL METRIC RANKS IN THE DIFFERENT RETINOBLASTOMA CATEGORY GRNS INFERRED FROM RETINOBLASTOMA MICROARRAY USING POSITIVE Z SCORE APPROACH BASED ON THE TOP 500 GENES IDENTIFIED BY EACH METRIC

	RB Blue Category Rank	RB Red Category Rank	RB Green Category Rank	Ranking Totals	Overall Rank
Degree Centrality	2nd	1st	1st	4	1st
Weighted Degree	2nd	1st	1st	4	1st
Eigenvector Centrality	1st	4th	1st	6	3rd
Combined Metric	5th	1st	1st	7	4th
Weighted Closeness	4th	6th	1st	11	5th
Weighted Betweenness	6th	5th	1st	12	6th

Weighted degree and degree centrality are the joint best performing metric, performing slightly better than the eigenvector centrality. The combined metric is the 4th best performing metric. Weighted closeness centrality, the 5th ranked metric, and weighted betweenness centrality, the 6th ranked metric, perform poorly compared to the other metrics.

7.8 Distribution of Genes in the Combined Metric Ranking Lists in the GRNs Inferred from the Retinoblastoma Microarray using the Positive Z Score Method

As noted in the retinoblastoma GRN category sections, none of the genes that appear in the top 20 ranked genes for the combined metric for any of the retinoblastoma categories appear in any other. This can be seen on the Venn diagram in figure B.21 in the appendix showing the distribution of genes in the top 20 combined metric rankings list in the categories. Furthermore, extending this to look at the top 100 combined metric rankings lists, shown in figure B.22 of the appendix, this is still the case, suggesting that this approach is suited to identifying specific genes associated with the different survival time

7.9 Network Level Metrics in the Retinoblastoma GRNs Inferred using the Negative Z Score Method

Having inferred networks for the three retinoblastoma categories using the positive z score method, the negative z score method can now be implemented to infer networks. As before, prior to looking at the individual genes in the networks, the network level metrics will be calculated and analysed. The same metrics will be used again; weighted degree assortativity, degree asssortativity, diameter, and network clustering. The scores for these are shown in table 7.23 below. The largest unique cliques will again be calculated for each of the networks for the two different datasets, and the genes that comprise these cliques will be investigated for both enrichment and the presence of retinoblastoma genes of interest using Génie.

TABLE 7.23 NETWORK LEVEL SCORES ACROSS THE CATEGORIES FOR THE RETINOBLASTOMA GRNS INFERRED USING THE NEGATIVE Z SCORE METHOD

Metric	RB Blue	RB Red	RB Green
Weighted Degree Assortativity	-0.242	-0.490	-0.579
Degree Assortativity	-0.242	-0.497	-0.600
Diameter	6.091	2.775	1.898
Network Clustering	0.424	0.621	0.599
Largest Clique Size	105	125	155

Looking at the network level scores, the value of the diameter once again is of interest. For both the datasets in the last chapter, and the networks constructed using the Z score positive – positive variant, the value of the diameter was greatest in the most advanced neuroblastoma category. This is also the case here. This again is an indication that the diameter is a network level property that is associated with disease stage, and more specifically, one could hypothesise that this could be due to genes involved in certain cellular processes not being able to easily communicate with each other, thereby contributing to the more advanced stage of disease. There is also a perfect correlation between the diameter ranking across the categories in the networks inferred using both the positive and negative variants of the Z score approach. In the next section, the size and composition of the largest unique cliques in the different categories will be looked at as this can be informative.

7.10 Largest Clique Calculation and Analysis in Retinoblastoma GRNs Inferred using Negative Z Score Method

The size and composition of the largest unique cliques in the different categories can be informative. It is also of interest to compare the largest unique cliques for the categories in the networks inferred from both datasets. A large number of common genes in the same category cliques across both datasets would be indicative of the same genes being involved in important cellular processes for specific retinoblastoma stages.

As well as identifying common and unique genes in the cliques for the different categories, it is informative to see whether these genes have been previously identified as being retinoblastoma genes of interest and appear on the list of Génie retinoblastoma genes. The list of genes that comprise the largest unique clique in each network category can also be checked for enrichment using the GenesetDB website. Table 7.24 below shows the size of the cliques in each category, whether any of the genes that comprise this clique are present in any other cliques, and finally whether any of the genes in the clique appear on the Génie retinoblastoma gene list.

TABLE 7.24 RETINOBLASTOMA GENES OF INTEREST IN THE LARGEST UNIQUE CLIQUES IN THE RETINOBLASTOMA GRNS INFERRED USING NEGATIVE Z SCORE METHOD

	Clique Size	Genes present in other RB cliques	Number of Génie genes and score
RB Blue	105	0	5, score of 1535
RB Red	125	0	1, score of 68

RB Green	155	0	16, score of 5005

Across the cliques, a number of retinoblastoma genes that appear on the list generated by Génie are present. As was the case with the largest unique cliques in the networks constructed using the positive Z score method, the RB Green category has both the highest number of Génie genes, and also the highest score of genes identified. As noted before, this might appear to be a surprising result, considering that the RB Green category network has been inferred from normal retinal samples. The sizes of the cliques in the three categories are more similar in size than was the case with the cliques for the categories in the networks inferred using the positive Z score method, with a range of 50, compared to a range of 142. It should also be noted that the RB Green clique is also the largest again. The RB Red category has the lowest number of genes identified, and the lowest score. A similar number of genes, 22 compared to 21, to the positive Z score method are present in the cliques, representing just 4.4 % of the complete list used. This lower proportion can be attributed to the more specific network inference method used; if both network construction variants are included, then 43 of the 500 genes are present in the cliques. The Venn diagram in figure B.23 of the appendix shows the distribution of the genes in the largest unique cliques in the networks inferred using the negative Z score method.

There are 385 genes in total across the three cliques. There are no genes common to either two or three cliques, with each gene only appearing in one clique. This result is indicative that the cellular processes are very different in the cliques in the different retinoblastoma categories. It would be expected that different genes are over-expressed in different retinoblastoma stages and healthy cellular processes; the distribution of genes in the cliques corresponds with this.

There are also no genes that are common between the largest unique cliques in the networks constructed using the positive Z score and negative Z score methods.

Having investigated the distribution of genes in the cliques, the next step is to investigate the enrichment of these genes using GenesetDB. For the genes that comprise the largest unique clique in the RB Blue category, 22 results are returned by GenesetDB. The top ten results returned are shown in table A.40 of the appendix. There are a number of results of interest here; sets relating to translation and mRNA splicing/processing may have relevance to retinoblastoma. Retinal photoreceptor cells are one of the most metabolically active cells in the body [141], and are reliant on high levels of gene transcription and translation for normal function. This characteristic may also facilitate tumour growth.

For the genes that comprise the largest unique clique in the RB Red category, only three results are returned from GenesetDB. This would imply that this combination of genes has not been identified as being involved in biological processes. The three results returned are shown in table A.41 of the appendix.

Unlike the previous two categories, that returned a small amount of enrichment results, the genes that comprise the largest unique clique in the RB Green category return a number of enrichment results. The top ten results returned are shown in table A.42 of the appendix. However, unlike with the enrichment results returned from the largest unique clique in the network inferred using the positive Z score method; none of these results are of interest.

7.11 RB Blue Category GRN Inferred using Negative Z Score Method

Table 7.25 shows the top 20 ranked genes for the combined metric in the RB Blue category GRN inferred using the negative Z score method. Two of these genes in the table appear in the list of 500 retinoblastoma genes generated by Génie; SP1, and SPAST. None of the genes that appear in the top 20 combine metric ranked list appear in any of the top 20 combined metric ranked lists for the other categories. As was the case with the largest unique cliques, this suggests clearly distinguishable differences between the retinoblastoma categories in terms of the genes that are playing important roles in cellular processes.

Despite 2 genes out of the 20 appearing on the Génie retinoblastoma list, and as was the case with the RB Blue network inferred using the positive Z score method, no enrichment results are returned for the 20 genes in the table below. This suggests that despite some of these genes being associated with retinoblastoma, they have not been identified as being involved together in biological processes.

TABLE 7.25 TOP 20 RANKED GENES FOR COMBINED METRIC IN RB BLUE GRN INFERRED USING NEGATIVE Z SCORE METHOD

Gene	*Combined Metric Rank*	*Gene*	*Combined Metric Rank*
SON	1	ZFR	11
SNAPC3	2	MRPL3	12
ACTR3	3	WAC	13
SP1	4	KIF5B	14
CSDE1	5	PPP1CB	15
KCMF1	6	ZNF532	16

MRPL50	7	HELZ	17
CNOT7	8	SPAST	18
TMEM33	9	ZNF138	19
MTMR4	10	MYST3	20

Table 7.26 below shows the Spearman Rank correlation scores for the rankings assigned to the genes by each metric in the RB Blue GRN inferred using the negative Z score method.

TABLE 7.26 SPEARMAN RANK CORRELATION SCORES OF RANKINGS ASSIGNED BY THE METRICS TO THE GENES IN RB BLUE GRN INFERRED USING THE NEGATIVE Z SCORE METHOD

	Degree Centrality	**Weighted Degree**	**Eigenvector Centrality**	**Closeness Centrality**	**Betweenness Centrality**
Degree Centrality	1	0.999	0.930	0.926	0.833
Weighted Degree	0.999	1	0.928	0.923	0.834
Eigenvector Centrality	0.930	0.928	1	0.967	0.659
Closeness Centrality	0.926	0.923	0.967	1	0.704
Betweenness Centrality	0.833	0.834	0.659	0.704	1

From the table above, it can be seen that all of metrics have a positive correlation greater than 0.65, and that a number of metrics have very strong correlations with other metrics. These strong correlations imply that there is a substantial overlap in the genes that the metrics rank highly, as was the case for a lot of the metrics in the last chapters. Therefore, we would expect that there will be quite a lot of overlap in the genes that these metrics identify.

Applying the scoring system to the top 20, top 100, and top 500 genes identified by each of metrics, the metrics can be ranked. The scores for the top 20 genes identified by each metric are shown in table 7.27 below.

TABLE 7.27 GÉNIE SCORES AND METRIC RANKING FOR THE TOP 20 RANKED GENES FOR EACH METRIC IN RB BLUE GRN INFERRED USING THE NEGATIVE Z SCORE METHOD

Metric	**Number of Génie retinoblastoma genes identified**	**Génie Score**	**Metric Rank**
Combined Metric	2	677	1st
Degree Centrality	1	428	2nd
Weighted Degree	1	428	2nd
Weighted Betweenness	1	428	2nd
Weighted Closeness	1	428	2nd
Eigenvector Centrality	0	0	6th

The combined metric performs best, identifying two genes, including the gene ranked 73rd for retinoblastoma by Génie, SP1. All of the other metrics apart from the eigenvector centrality also identify Génie; SP1. Applying the scoring system to the top 100 ranked genes for each metric, the results in table 7.28 below are given.

TABLE 7.28 GÉNIE SCORES AND METRIC RANKING FOR THE TOP 100 RANKED GENES FOR EACH METRIC IN RB BLUE GRN INFERRED USING THE NEGATIVE Z SCORE METHOD

Metric	**Number of Génie retinoblastoma genes identified**	**Génie Score**	**Metric Rank**
Weighted Betweenness	5	1644	1st

Weighted Closeness	4	1425	2nd
Combined Metric	3	1156	3rd
Degree Centrality	2	677	4th
Weighted Degree	2	677	4th
Eigenvector Centrality	2	677	4th

Applying the scoring system to the top 100 genes identified by the metrics, the weighted betweenness centrality is the best performing metric, identifying 5 genes. Finally, extending the scoring system to the top 500 genes, the following results shown in table 7.29 below are given.

TABLE 7.29 GÉNIE SCORES AND METRIC RANKING FOR THE TOP 500 RANKED GENES FOR EACH METRIC IN RB BLUE GRN INFERRED USING THE NEGATIVE Z SCORE METHOD

Metric	**Number of Génie retinoblastoma genes identified**	**Génie Score**	**Metric Rank**
Combined Metric	25	5968	1st
Weighted Closeness	21	5358	2nd
Degree Centrality	22	5252	3rd
Weighted Degree	22	5252	3rd
Eigenvector Centrality	22	5161	5th
Weighted Betweenness	22	4807	6th

The combined metric performs best, followed by weighted closeness and then degree and weighted degree centrality. Three of the top 25 ranked genes by Génie for retinoblastoma are identified by the combined metric; CDK4 ranked 15th, BRAF ranked 22nd, and CDKN1B ranked 23rd. Extending this, 5 of the top 50 ranked genes are identified. Across the whole network, 76 out of the 500 retinoblastoma genes generated by Génie are present; compared to 53 that are present in the RB Blue network inferred using the positive Z score method. This result might indicate that under-expression of genes is more informative in the RB Blue category than gene amplification; based on more Génie retinoblastoma genes of interest being present in the network inferred using the negative Z score method.

7.12 RB Red Category GRN Inferred using Negative Z Score Method

Table 7.30 shows the top 20 ranked genes for the combined metric in the RB Red category GRN inferred using the negative Z score method. None of the genes in the table appear in the list of 500 retinoblastoma genes generated by Génie. Also, none of the genes that appear in the top 20 combined metric ranked list appear in any of the top 20 combined metric ranked lists for the other categories. As was the case with the previous category, this suggests that different genes are playing important roles in the different retinoblastoma stages.

No enrichment results are returned for the list of top 20 ranked genes, implying that this particular combination of genes has not been identified as being involved in any significant biological processes. None of the genes being present in the top 500 list returned by Génie also suggests that this list of genes might not be as biologically relevant as the list of 20 genes returned by the previous retinoblastoma category.

TABLE 7.30 TOP 20 RANKED GENES FOR COMBINED METRIC IN RB RED GRN INFERRED USING NEGATIVE Z SCORE METHOD

Gene	*Combined Metric Rank*	*Gene*	*Combined Metric Rank*
FXYD7	1	LEFTY1	11
RBMS1	2	C6ORF15	12
GPR37L1	3	MADCAM1	13
IL28B	4	GPR149	14
BARHL2	5	CTF1	15
DNM1P35	6	C9ORF141	16
OR6C76	7	MCCD1	17
DRD5	8	IGFBP1	18
FAM115C	9	ADAMTS7	19
PLA2G2D	10	DKFZP779M0652	20

Table 7.31 below shows the Spearman Rank correlation scores for the rankings assigned to the genes by each metric in the RB Red GRN inferred using the negative Z score method.

TABLE 7.31 SPEARMAN RANK CORRELATION SCORES OF RANKINGS ASSIGNED BY THE METRICS TO THE GENES IN RB RED GRN INFERRED USING NEGATIVE Z SCORE METHOD

	Degree Centrality	**Weighted Degree**	**Eigenvector Centrality**	**Closeness Centrality**	**Betweenness Centrality**
Degree Centrality	1	0.999	0.998	0.998	0.912
Weighted Degree	0.999	1	0.997	0.997	0.913

Eigenvector Centrality	0.998	0.997	1	0.996	0.899
Closeness Centrality	0.998	0.997	0.996	1	0.913
Betweenness Centrality	0.912	0.913	0.899	0.913	1

There are extremely strong correlation scores between the metrics in this category, as can be seen in table above, with the weakest correlation of 0.899. These extremely strong correlations imply that there will be a great deal of overlap in the genes that the metrics rank highly, as was the case for a lot of the metrics in the last chapters.

Applying the scoring system to the top 20, top 100, and top 500 genes identified by each of metrics, the metrics can be ranked. The scores for the top 20 genes identified by each metric are shown in table 7.32 below.

TABLE 7.32 GÉNIE SCORES AND METRIC RANKING FOR THE TOP 20 RANKED GENES FOR EACH METRIC IN RB RED GRN INFERRED USING THE NEGATIVE Z SCORE METHOD

Metric	**Number of Génie retinoblastoma genes identified**	**Génie Score**	**Metric Rank**
Degree Centrality	0	0	1st
Weighted Degree	0	0	1st
Weighted Betweenness	0	0	1st
Weighted Closeness	0	0	1st
Eigenvector Centrality	0	0	1st
Combined Metric	0	0	1st

None of the metrics identify a single Génie retinoblastoma gene of interest. Applying the scoring system to the top 100 ranked genes for each metric, the results in table 7.33 below are given.

TABLE 7.33 GÉNIE SCORES AND METRIC RANKING FOR THE TOP 100 RANKED GENES FOR EACH METRIC IN RB RED GRN INFERRED USING THE NEGATIVE Z SCORE METHOD

Metric	**Number of Génie retinoblastoma genes identified**	**Génie Score**	**Metric Rank**
Weighted Betweenness	2	217	1st
Combined Metric	2	217	1st
Degree Centrality	1	68	3rd
Weighted Degree	1	68	3rd
Weighted Closeness	1	68	3rd
Eigenvector Centrality	1	68	3rd

The weighted betweenness and the combined metric are the joint best performing metrics, both identifying the same two genes. All of the other metrics identify one of these two genes. All of the metrics perform badly at identifying genes, although the performance is similar to that of the metrics in the RB Red category in the network inferred using the positive Z score method. Finally, extending the scoring system to the top 500 genes, the following results shown in table 7.34 below are given.

TABLE 7.34 GÉNIE SCORES AND METRIC RANKING FOR THE TOP 500 RANKED GENES FOR EACH METRIC IN RB RED GRN INFERRED USING THE NEGATIVE Z SCORE METHOD

Metric	Number of Génie retinoblastoma genes identified	Génie Score	Metric Rank
Degree Centrality	8	1172	1st
Weighted Degree	8	1172	1st
Weighted Betweenness	8	1172	1st
Weighted Closeness	8	1172	1st
Eigenvector Centrality	8	1172	1st
Combined Metric	8	1172	1st

All of the metrics identify the same eight genes, which are also all of the Génie retinoblastoma genes of interest that are present in the whole network. There are no genes in either the top 25 or top 50 ranking by Génie of retinoblastoma genes. Compared to the RB Red network inferred using the positive Z score method, there are 45 fewer Génie retinoblastoma genes present. In contrast to the RB Blue category, this could indicate that over-expression of genes is more informative in the RB Red category than gene under-expression; based on more Génie retinoblastoma genes of interest being present in the network inferred using the positive Z score method.

7.13 RB Green Category GRN Inferred using Negative Z Score Method

Table 7.35 shows the top 20 ranked genes for the combined metric in the RB Green category GRN inferred using the negative Z score method. Three of the genes in the table, E2F1 ranked 5th by Génie, PLK1 ranked 220th by Génie, and TP73 ranked 41st by Génie, appear in the list of retinoblastoma genes of interest. None of the genes that appear in the top 20 combined metric ranked list appear in any of the top 20 combined metric ranked lists for the other categories. This again suggests that different genes are playing important roles in the different retinoblastoma stages.

Table A.43 of the appendix shows the top twenty enrichment results for this list of genes. The top twenty results are shown this time as whilst none of the top ten results are directly involved in retinal or retinoblastoma biology, the 16th result returned concerns the regulation of the previously mentioned E2F1. The 17th result returns also details increased incidence of tumours. The genes PLK1, POLA2, and TP73 that are in the list of top 20 ranked genes, have been identified as being involved in the regulation of the gene E2F1.

TABLE 7.35 TOP 20 RANKED GENES FOR COMBINED METRIC IN RB GREEN GRN INFERRED USING NEGATIVE Z SCORE METHOD

Gene	*Combined Metric Rank*	*Gene*	*Combined Metric Rank*
CHAF1A	1	TROAP	11
WDR34	2	CPXM1	12
ZWINT	3	POLA2	13
E2F1	4	KIF18B	14
REC8	5	DKFZP434L187	15
LRDD	6	MCM2	16
GLT25D1	7	FEN1	17
CENPM	8	PLK1	18
RCC2	9	CDCA2	19
STMN1	10	TP73	20

Table 7.36 below shows the Spearman Rank correlation scores for the rankings assigned to the genes by each metric in the RB Green GRN inferred using the negative Z score method.

TABLE 7.36 SPEARMAN RANK CORRELATION SCORES OF RANKINGS ASSIGNED BY THE METRICS TO THE GENES IN RB GREEN GRN INFERRED USING NEGATIVE Z SCORE METHOD

	Degree Centrality	**Weighted Degree**	**Eigenvector Centrality**	**Closeness Centrality**	**Betweenness Centrality**
Degree Centrality	1	0.999	0.999	0.992	0.806
Weighted Degree	0.999	1	0.999	0.992	0.805
Eigenvector Centrality	0.999	0.999	1	0.991	0.804
Closeness Centrality	0.992	0.992	0.991	1	0.804
Betweenness Centrality	0.806	0.805	0.804	0.804	1

Once again, there are strong correlation scores between the metrics, with the minimum correlation score being 0.80. Again, this implies significant overlap in the genes ranked highly by the different metrics. Applying the scoring system to the top 20, top 100, and top 500 genes identified by each of metrics, the metrics can be ranked. The scores for the top 20 genes identified by each metric are shown in table 7.37 below.

TABLE 7.37 GÉNIE SCORES AND METRIC RANKING FOR THE TOP 20 RANKED GENES FOR EACH METRIC IN RB GREEN GRN INFERRED USING THE NEGATIVE Z SCORE METHOD

Metric	**Number of Génie retinoblastoma genes identified**	**Génie Score**	**Metric Rank**
Weighted Betweenness	3	1237	1st
Combined Metric	3	1237	1st
Degree Centrality	2	777	3rd
Weighted Degree	2	777	3rd
Weighted Closeness	2	777	3rd
Eigenvector Centrality	2	777	3rd

The weighted betweenness and the combined metric are the joint best performing metrics, both identifying the same three genes, including TP73. The other metrics all identify the genes E2F1 and PLK1 that the weighted betweenness and combined metric also identify. Applying the scoring system to the top 100 ranked genes for each metric, the results in table 7.38 below are given.

TABLE 7.38 GÉNIE SCORES AND METRIC RANKING FOR THE TOP 100 RANKED GENES FOR EACH METRIC IN RB GREEN GRN INFERRED USING THE NEGATIVE Z SCORE METHOD

Metric	**Number of Génie retinoblastoma genes identified**	**Génie Score**	**Metric Rank**
Degree Centrality	14	4520	1st
Weighted Degree	14	4520	1st
Weighted Closeness	14	4520	1st
Eigenvector Centrality	14	4520	1st

Combined Metric	13	4028	5th
Weighted Betweenness	10	2644	6th

The degree, weighted degree, weighted closeness and eigenvector centrality all identify the same 14 genes. As noted, the strong correlation scores between the metrics means that we might expect the metrics to identify the same genes. Three of the top 50 ranked genes for retinoblastoma; E2F1, BRCA1, and TP73, are present amongst these 14 genes. Finally, extending the scoring system to the top 500 genes, the following results shown in table 7.39 below are given.

TABLE 7.39 GÉNIE SCORES AND METRIC RANKING FOR THE TOP 500 RANKED GENES FOR EACH METRIC IN RB GREEN GRN INFERRED USING THE NEGATIVE Z SCORE METHOD

Metric	**Number of Génie retinoblastoma genes identified**	**Génie Score**	**Metric Rank**
Weighted Closeness	45	12866	1st
Combined Metric	45	12866	1st
Degree Centrality	44	12731	3rd
Weighted Degree	44	12731	3rd
Weighted Betweenness	44	12731	3rd
Eigenvector Centrality	44	12731	3rd

The combined metric and weighted closeness are the best performing metrics, identifying the same 45 genes. These two just outperform the other metrics that identify 44 genes. 7 of the 45 genes that the combined metric and weighted closeness identify are in the top 50 Génie ranked genes for retinoblastoma. There are 49 retinoblastoma genes of interest present in this network. Again, it might be surprising for so many retinoblastoma genes to be identified in the network inferred from healthy samples, as before, there are a variety of reasons as to why this is the case, and again highlights the role of clinicians and biologists to infer biological meaning from these results. 27 Génie retinoblastoma genes of interest are present in the RB Green network inferred using the positive Z score method, indicating that under-expression of genes is more informative in the RB Green category than gene amplification; based on more Génie retinoblastoma genes of interest being present in the network inferred using the negative Z score method. Another interpretation of this is that these genes might be involved in retinoblastoma cellular processes when they are significantly over-expressed, thereby explaining their presence in the network inferred from healthy retinal samples using the negative Z score method.

7.14 Metric Performance in the GRNs Inferred from the Retinoblastoma Microarray using the Negative Z Score Method

Previously, each metric was ranked based on the top 20, top 100 and top 500 genes they identified from each retinoblastoma category GRN inferred using the negative Z score method. This ranking is based on the score of the identified genes in the Génie retinoblastoma gene list.

The results of these metric rankings can now be presented together, allowing an analysis of which metric is the best performing overall. Assigning an equal weighting to the scores of the top 20, top 100, and top 500 ranked genes for each metric in each category, the following ranks shown in table 7.40 below are given.

TABLE 7.40 OVERALL METRIC RANKS IN THE DIFFERENT RETINOBLASTOMA CATEGORY GRNS INFERRED FROM RETINOBLASTOMA MICROARRAY USING NEGATIVE Z SCORE APPROACH

	RB Blue Category Rank	RB Red Category Rank	RB Green Category Rank	Ranking Totals	Overall Rank
Combined Metric	1st	1st	2nd	4	1st
Weighted Closeness	2nd	3rd	1st	6	2nd
Degree Centrality	3rd	3rd	2nd	8	3rd
Weighted Degree	3rd	3rd	2nd	8	3rd
Weighted Betweenness	3rd	1st	6th	10	5th
Eigenvector Centrality	6th	3rd	2nd	11	6th

The combined metric is the best performing metric overall, followed by weighted closeness. If the rankings for just the top 20 genes in each category are taken into consideration, the following scores shown in table 7.41 are obtained.

TABLE 7.41 OVERALL METRIC RANKS IN THE DIFFERENT RETINOBLASTOMA CATEGORY GRNS INFERRED FROM RETINOBLASTOMA MICROARRAY USING NEGATIVE Z SCORE APPROACH BASED ON THE TOP 20 GENES IDENTIFIED BY EACH METRIC

	RB Blue Category Rank	RB Red Category Rank	RB Green Category Rank	Ranking Totals	Overall Rank
Combined Metric	1st	1st	1st	3	1st
Weighted Betweenness	2nd	1st	1st	4	2nd

Degree Centrality	2nd	1st	3rd	6	3rd
Weighted Degree	2nd	1st	3rd	6	3rd
Weighted Closeness	2nd	1st	3rd	6	3rd
Eigenvector Centrality	6th	1st	3rd	10	6th

The combined metric is again the best performing metric, followed by weighted betweenness. The eigenvector centrality is the worst performing metric again. If we take the ranking for the top 100 genes into account, the following results shown in table 7.42 are obtained.

TABLE 7.42 OVERALL METRIC RANKS IN THE DIFFERENT RETINOBLASTOMA CATEGORY GRNS INFERRED FROM RETINOBLASTOMA MICROARRAY USING NEGATIVE Z SCORE APPROACH BASED ON THE TOP 100 GENES IDENTIFIED BY EACH METRIC

	RB Blue Category Rank	RB Red Category Rank	RB Green Category Rank	Ranking Totals	Overall Rank
Weighted Closeness	2nd	3rd	1st	6	1st
Degree Centrality	4th	3rd	1st	8	2nd
Weighted Degree	4th	3rd	1st	8	2nd
Weighted Betweenness	1st	1st	6th	8	2nd
Eigenvector Centrality	4th	3rd	1st	8	2nd

Combined Metric	3rd	1st	5th	9	6th

Weighted closeness is the Degree centrality is the best performing metric, followed by four of the other metrics that all exhibit the same performance. The combined metric, which was the best performing metric the previous two times, is the worst performing metric this time. Finally, taking the top 500 gene into account, the following results in table 7.43 are obtained.

TABLE 7.43 OVERALL METRIC RANKS IN THE DIFFERENT RETINOBLASTOMA CATEGORY GRNS INFERRED FROM RETINOBLASTOMA MICROARRAY USING NEGATIVE Z SCORE APPROACH BASED ON THE TOP 500 GENES IDENTIFIED BY EACH METRIC

	RB Blue Category Rank	**RB Red Category Rank**	**RB Green Category Rank**	**Ranking Totals**	**Overall Rank**
Combined Metric	1st	1st	1st	3	1st
Weighted Closeness	2nd	1st	1st	4	2nd
Degree Centrality	3rd	1st	3rd	7	3rd
Weighted Degree	3rd	1st	3rd	7	3rd
Eigenvector Centrality	5th	1st	3rd	9	5th
Weighted Betweenness	6th	1st	3rd	10	6th

This time the combined metric just out-performs weighted closeness centrality, with these two metrics showing similar performance. Eigenvector centrality, the 5th ranked metric, and weighted betweenness centrality, the 6th ranked metric, perform poorly compared to the other metrics.

7.15 Distribution of Genes in the Combined Metric Ranking Lists in the GRNs Inferred from the Retinoblastoma Microarray using the Negative Z Score Method

As noted in the retinoblastoma category sections for the networks inferred using the positive Z score method, none of the genes that appear in the top 20 ranked list for the combined metric for any category appear in any other, shown in figure B.21 of the appendix. This is also the case for the networks inferred using the negative Z score method, shown in figure B.24 of the appendix. As well as this, no top 20 ranked genes for the combined metric for any category in the network inferred using the positive Z score method are present in any of the top 20 combined metric ranking genes for the GRNs inferred using the negative Z score method, and vice-versa. This demonstrates that clearly distinguishable genes are involved in the different networks constructed using the two different methods; it is not a case of certain genes that are identified as being high ranking in the network for one category using one of the methods, is then identified as being high ranking in the network for a different category using the other method. This would suggest that implementing the two different methods based on the nature of the correlations allows identification of very specific genes, which may be beneficial to clinicians and biologists looking to identify particular genes being important in particular diseases categories. Extending this to look at the top 100 ranked genes in the combined metric rankings list, shown in figure B.25 of the appendix, this is also the case.

7.16 Discussion

In this chapter, a modified version of the Z score network inference method used in the previous two chapters was introduced. The principal reason for this was to enable a specific means to infer separate networks involving gene over expression and gene under expression, as oppose to the absolute correlation approach adopted before. As a result of implementing this approach, unique and specific genes were identified as being important in different categories of a proprietary retinoblastoma dataset based on both gene over expression, and gene under expression. This can be seen by the fact that there are no common genes either between the top 100 combined metric ranked genes for each category, either between the categories inferred using the same network inference method, and also between the categories of the networks inferred using the two different network inference methods. This is also the case with the genes that comprise the largest unique cliques of the networks inferred.

In the networks inferred using the positive Z score method, there are a number of perhaps unexpected findings. These mostly concern the genes that comprise the largest unique cliques in the GRNs, and the enrichment results these yield. As noted, it might be surprising to find more genes identified as being amongst the 500 most important for retinoblastoma by Génie in the largest unique clique for the RB Green category, than in the largest unique cliques for the other two categories. The perhaps paradoxical enrichment results yielded for this clique; results pertaining to both healthy retinal cell processes and diseased retinal processes might also be a result that at first has little sense, although, can be attributed to the enrichment results not distinguishing between gene over expression and under expression. However, looking at the top ranked genes in each of the three networks and the genes that make up these networks, there are a number of expected results. The RB Blue Category, the most advanced retinoblastoma stage, has a greater number of genes identified by Génie in the top 20 ranked genes than the other two networks. Furthermore, both the RB Blue, 48 genes, and the RB Red, 53 genes, have more of these top 500 retinoblastoma genes of interest in their networks, than the RB Green network has, 27 genes. We would expect that more genes previously identified as being important for retinoblastoma would be present in the networks inferred from the retinoblastoma samples than the network inferred from healthy retinal samples.

A similar result with the composition of the largest unique cliques occurs in the networks inferred using the negative Z score method. There are again more genes identified as being amongst the 500 most important for retinoblastoma by Génie in the largest unique clique for the RB Green category, than in the largest unique cliques for the other two categories. No enrichment results are yielded this time that relate either to retinoblastoma or the retina in general for any of the cliques. In the GRNs, there is one particular enrichment result that is yielded for the RB Green category, relating to the regulation of the gene E2F1, ranked 5th by Génie for retinoblastoma. pRB, the protein that is encoded by the RB1 gene, is directly responsible for blocking the expression of E2F1[142]. Therefore, the expression of E2F1 is lower in the RB Green samples that have a normal RB1 gene, than in the RB Red and RB Blue samples. E2F1 is one of the top 20 ranked genes in the RB Green network, suggesting therefore that the gene regulatory processes responsible for under expressing E2F1 might be a property that distinguishes healthy retina from retinoblastoma. There are also more retinoblastoma genes of interest in the top 20 ranked genes for the RB Green network, than for the RB Blue network, 2, and the RB Red network, 0. Clear differences can be seen between the two networks inferred from the retinoblastoma samples; there are 76 retinoblastoma genes of interest in the RB Blue network, whilst only 8 in the RB Red network. There are 49 retinoblastoma genes of interest in the RB Green network.

The use of the modified implementation of the Z score network inference method is promising; unique genes in all categories of the retinoblastoma dataset were identified, despite the small number of samples. A number of previously identified retinoblastoma genes of interest were present amongst these genes, although the need for expert interpretation from specialists such as biologists and clinicians was shown by the network inferred from the healthy retinal samples containing a number of genes identified by Génie. This shows that despite the promising results obtained, there is still a need for manual verification from experts.

7.17 Conclusions

To conclude, in this chapter a refinement of the novel inference method was introduced, to specifically distinguish between genes whose expression is amplified together, and genes whose expression is under expressed together. This method was applied to a retinoblastoma microarray dataset from samples collected at BCH to infer three categories of GRN. Unlike the previous microarray datasets used in the work, a normal reference GRN was amongst the categories of network inferred, due to the presence of what appeared to be normal tissues amongst the samples in the dataset. This allowed comparisons to be made against this normal GRN. Again, graph theory metrics were used to identify high ranking genes, and the metrics were scored and compared based on their ability to identify genes identified by the text-mining tool Génie as being important for retinoblastoma. The next chapter details approaches for making the GRN inference methods and tools for calculating the metrics used in this work available to clinicians and biologists, as well as the wider scientific community. This is done to improve the accessibility and usability of the methods.

Chapter 8

Improving the Accessibility of GRN Inference and Analysis Tools

8.1 Introduction

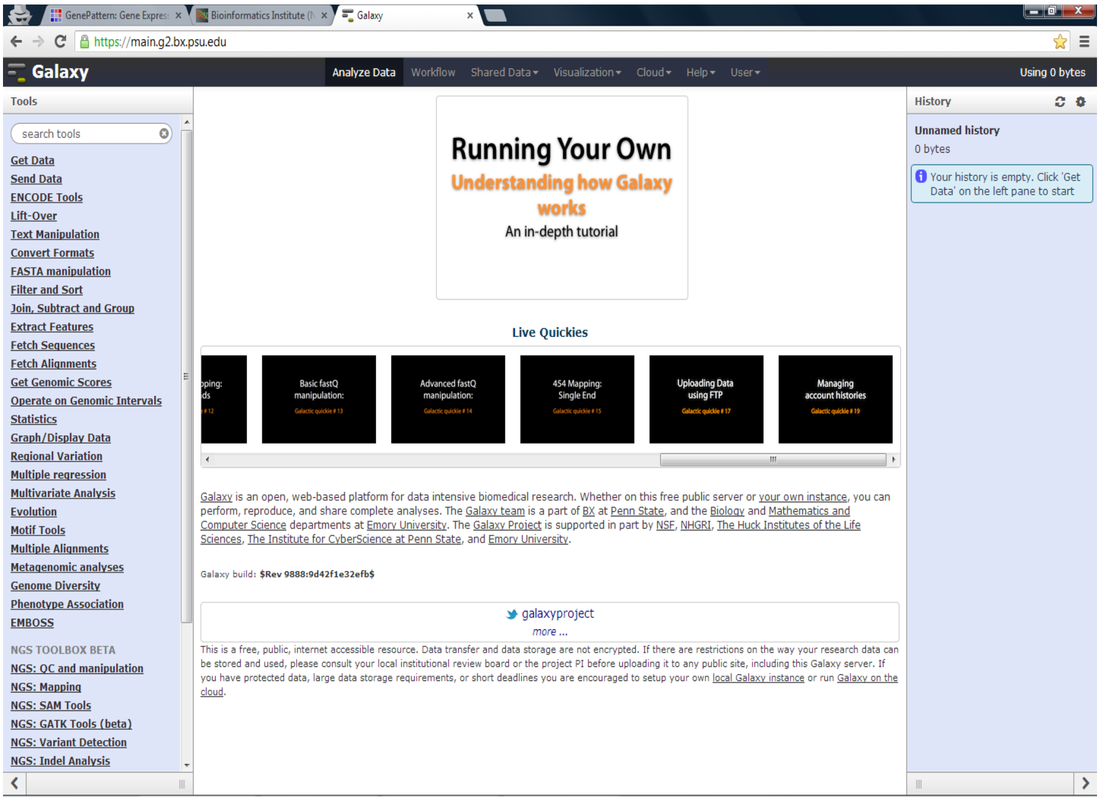
In this chapter, a number of options for making the network inference methods used in this work more accessible are presented. The work in this chapter addresses the accessibility and usability outlined in the first objective. Having specifically seen in chapters 5, 6, and 7 how the application of the novel network inference method identifies unique genes of interest specifically associated with disease stages, one important area to consider is how easily it can be adopted by clinicians and biologists to interrogate microarray data. To date, the approach adopted in this work has been to infer and then analyse GRNs using R from microarray datasets that have been flagged as being of interest by clinicians and biologists. Following feedback from clinicians and biologists, they would like to be able to use the network inference method themselves on microarray datasets, but without having to learn a programming language.

Specialist programming skills required for network inference methods in languages such as R, Matlab, and Python, are proving a barrier to widespread use of network inference approaches to microarray data by biologists and clinicians [9]. There is also great variation in the programming languages that different research groups involved in microarray analysis use, posing potential problems to collaborative research. One drawback of this is that the potential benefit of greater use of network inference techniques to analyse microarray data is not being fully harnessed, especially considering that biologists and clinicians are those with the expert knowledge that are the best-equipped to interpret the results. To counter this, the Z score network inference method should be accessible to users that do not possess these specialist programming skills. Therefore, the implementation of the Z score network inference method, along with other GRN inference and analysis tools for microarray data, on a web based interface is one of the options presented. This was originally going to be presented as one of the recommendations for the future for building upon this work, however following discussions with biologists and clinicians, the immediate need for an easy means to access network inference tools for microarray analysis became apparent.

Despite there being a number of web based platforms that host bioinformatics tools, there is a lack of network inference and analysis tools available on these platforms. A number of these web based platforms, such as GenePattern [143] and Genevestigator [144], host gene expression analysis tools but these fall mainly into the differential analysis category. There are no tools to infer GRNs from gene expression data. Therefore, following advice from researchers at the Bioinformatics Institute at the University of Auckland, including Director of the New Zealand Bioinformatics Institute Associate Professor Cristin Print, a number of GRN inference and analysis tools have been developed using the Galaxy framework [10]; an open, web-based platform for data intensive biomedical research. The package RGalaxy [145] was used to enable the R based scripts that the Z score network inference was coded in to be accessed via the Galaxy interface, as well as the other network inference and analysis tools.

For the purposed of this project, a personalised local host version of Galaxy running on a dedicated Ubuntu machine that can be remotely accessed has been set up. This has been done for two reasons; firstly to evaluate the feasibility of such a tool before a publically accessible version is made available, and secondly due to the confidential nature of a number of unpublished datasets that the biologists and clinicians would not feel comfortable analysing using a publically accessible tool. It is envisioned that this local host version will be retained in the future alongside any publicly accessible version, specifically for use with confidential data. As well as the novel network inference method used in chapters 5 and 6 in this work, a number of other network inference and analysis tools have also been made available. These include the WGCNA network inference method used in chapter 4, network metric calculation tools for analysing existing networks, enrichment analysis tools, and tools to export subnets of interest to the network visualisation software Cytoscape [146]. There are a number of tools on the public Galaxy server that are not of use, and have been removed to make the interface a lot simpler. On the screenshot below, these tools can be seen on the left-hand side, as well as the welcome screen.

FIGURE 8.1 PUBLIC GALAXY SERVER WELCOME SCREEN



Your history is empty. Click 'Get Data' on the left pane to start

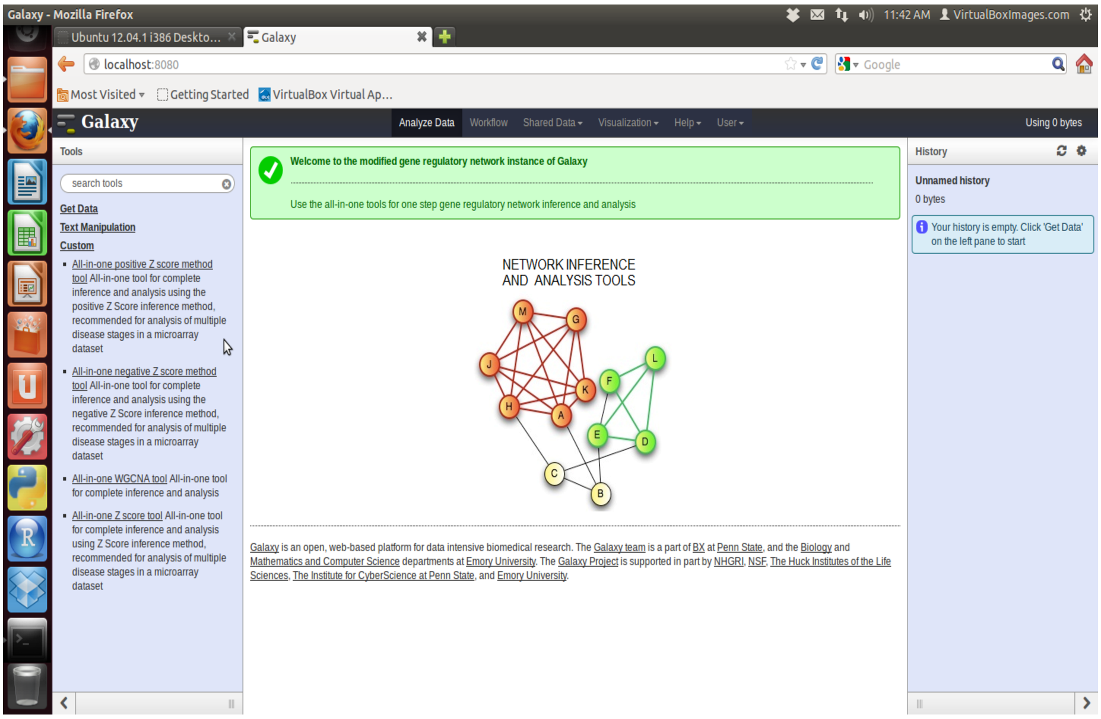
Galaxy is an open, web-based platform for data intensive biomedical research. Whether on this free public server or your own instance, you can perform, reproduce, and share complete analyses. The Galaxy team is a part of BX at Penn State, and the Biology and Mathematics and Computer Science departments at Emory University. The Galaxy Project is supported in part by NSF, NHGRI, The Huck Institutes of the Life Sciences, The Institute for CyberScience at Penn State, and Emory University.

Galaxy build: **$Rev 9888:9d42f1e32efb$**

This is a free, public, internet accessible resource. Data transfer and data storage are not encrypted. If there are restrictions on the way your research data can be stored and used, please consult your local institutional review board or the project PI before uploading it to any public site, including this Galaxy server. If you have protected data, large data storage requirements, or short deadlines you are encouraged to setup your own local Galaxy instance or run Galaxy on the cloud.

In order to facilitate use of the network inference and analysis tools using Galaxy, as well as removing the non-required tools, the welcome screen has also been simplified. The screenshot below shows the welcome screen of the local personalised Galaxy version.

FIGURE 8.2 PERSONALISED LOCAL HOST VERSION OF GALAXY WELCOME SCREEN



Get Data

Text Manipulation

Custom

- All-in-one positive Z score method tool All-in-one tool for complete inference and analysis using the positive Z Score inference method, recommended for analysis of multiple disease stages in a microarray dataset
- All-in-one negative Z score method tool All-in-one tool for complete inference and analysis using the negative Z Score inference method, recommended for analysis of multiple disease stages in a microarray dataset
- All-in-one WGCNA tool All-in-one tool for complete inference and analysis
- All-in-one Z score tool All-in-one tool for complete inference and analysis using Z Score inference method, recommended for analysis of multiple disease stages in a microarray dataset

Welcome to the modified gene regulatory network instance of Galaxy

Use the all-in-one tools for one step gene regulatory network inference and analysis

Galaxy is an open, web-based platform for data intensive biomedical research. The Galaxy team is a part of BX at Penn State, and the Biology and Mathematics and Computer Science departments at Emory University. The Galaxy Project is supported in part by NHGRI, NSF, The Huck Institutes of the Life Sciences, The Institute for CyberScience at Penn State, and Emory University.

Unnamed history

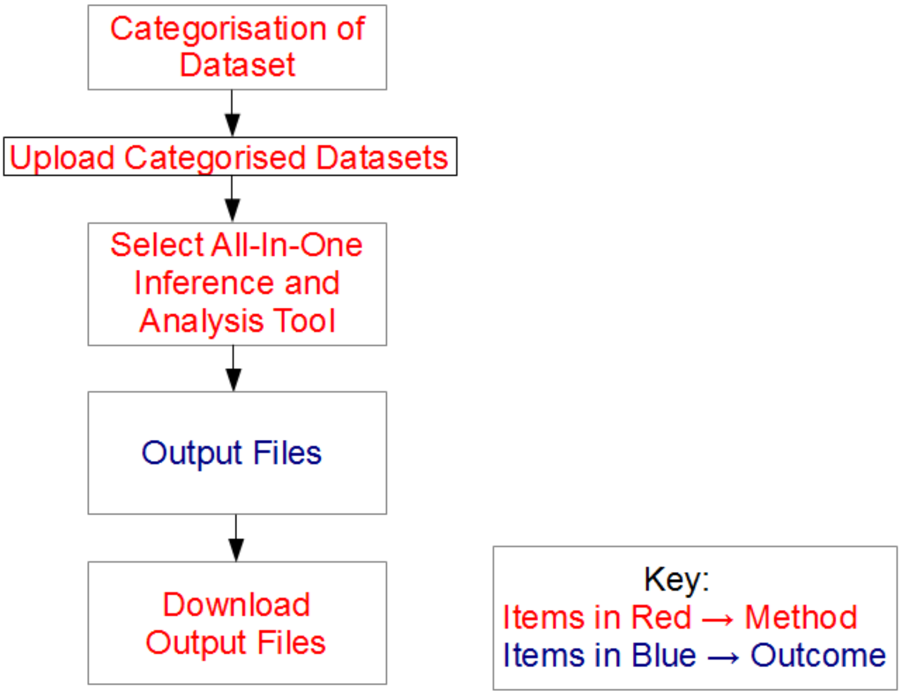
0 bytes

Your history is empty. Click 'Get Data' on the left pane to start

As can be seen, there is a much cleaner and simpler interface, with only three categories of tools. The Get Data and Text Manipulation tool sections have been maintained, and the network inference and analysis tools are found under the custom section. A number of additional screenshots are included in appendix C.

The flowchart on the next page shows an overview of the process for the biologists and clinicians to analyse microarray datasets using the local host version of Galaxy. All that is required is for the end user to categorise the datasets based on their own criteria, and upload them to the platform. The all-in-one tool then carries out all the processes seen in the flowchart in figure 3.2 that required specialist programming skills, and presents the output for the end user to download for their own further analysis and interpretation. Two main benefits have been achieved; the need for an end user with specialist programming skills has been removed; and the need for this end user to go back and forth to the clinicians and biologists for guidance on the classification criteria and expert interpretation of the results has also been removed. Figure 8.3 below shows the new simplified process.

FIGURE 8.3 OVERVIEW OF PROCESS FOR NETWORK INFERENCE AND ANALYSIS USING PERSONALISED GALAXY LOCAL HOST SERVER


Categorisation of Dataset
Upload Categorised Datasets
Select All-In-One Inference and Analysis Tool
Output Files
Download Output Files
Key:
Items in Red → Method
Items in Blue → Outcome

8.2 Making Tools Accessible to the Wider Scientific Community

As well as the local host implementation of Galaxy aimed at clinicians and biologists, there are alternative options that can make the tools implemented available to the wider scientific community. These fall into two categories; without hosting, and with hosting. The first of these is to publish the Galaxy tools developed directly to the Galaxy Tool Shed, found at http://toolshed.g2.bx.psu.edu/. This allows other users with personalised Galaxy installations to directly implement these, thus allowing all users of their respective installation to access these tools. This option is aimed at users responsible for installing and maintaining Galaxy installations, rather than at clinicians and biologists. In addition, an R package can be created and contributed to CRAN, http://cran.r-project.org/, allowing users with knowledge of R to access and use the tools, and R scripts can be made available online, at sites such as http://sourceforge.net/.

The second means by which to make the tools accessible to the wider scientific community is to host a publically accessible version of Galaxy on a dedicated separate server. The main benefit of this is that a publically accessible server specifically tailored to microarray analysis can be deployed. Additionally, access will not be limited to biologists and clinicians affiliated to a particular research group with a local version of Galaxy. However, in order to provide this an extensive validation and verification process will be required to ensure the robustness of both the tools implemented and the infrastructure used, and extensive hardware resources will be required. To illustrate this, one instance of a publically accessible Galaxy server consists of a Dell blade/m1000e chassis, with 128 cores, 1TB RAM, and 72TB raw array storage.

8.3 Summary

The first section of this chapter detailed the implementation of GRN inference and analysis tools on a local host of the Galaxy biomedical analysis server. This has been presented as a means to improve the accessibility and usability of GRN inference and analysis tools to end users without specialist programming skills, and will require extensive evaluation and testing before they can be made available on a publically available web-based server. Initial feedback from researchers who have used the local version based at the Bioinformatics Institute at the

University of Auckland has been promising, suggesting that this is potential means by which to improve the accessibility and usability of GRN inference and analysis tools for microarray datasets. The second section provided an overview of ways that the tools can be made available to the wider scientific community, which includes making the tools available on the Galaxy Tool Shed. In the next chapter, the overall conclusions for the thesis are presented, along with limitations and areas for future development.

Chapter 9

Conclusions

The aim of this thesis was to investigate the progression and evolution of tumours, by inferring and analysing GRNs for different evolutionary and clinical stages of cancer microarray datasets. It has been shown in this thesis that applying graph theory metrics to GRNs inferred using a novel inference method achieved this aim, and identified a number of genes with specific clinical or evolutionary stages.

Inferring GRNs from microarray datasets to investigate the progression and evolution of tumours is a complex process. Three specific objectives were set out at the beginning of work in order to achieve the aim of the work; contributions towards these three objectives are outlined below.

9.1 Contributions

9.1.1 Development of a Novel Network Inference Method

The first objective of the work was to develop a network inference method specifically designed to infer a number of GRNs for a number of different evolutionary or disease stage categories from a single microarray dataset. Following feedback from biologists and clinicians, a novel network inference method was developed. Comparing the scores of the top ranked genes identified by the node level metrics in the five glioblastoma survival category GRNs inferred using this method, to the top ranked genes in the equivalent networks inferred using WGCNA, this method consistently identified higher scoring glioblastoma genes across all of the survival stage networks.

The application of this network inference method to other microarray datasets, such as neuroblastoma and retinoblastoma, also identified unique genes with significantly varying gene expression in each disease stage. This was something that was not the case when the WGCNA method was used to infer GRNs for the different survival categories within the glioblastoma dataset, and was something that the biologists felt went against biological principles. The identification of unique genes with varying gene expression levels by the novel method is both in concordance with the underlying biology, and also gave greater confidence to the biologists that this approach was something that could provide useful results for them. A further refined version of the novel inference method was applied to a retinoblastoma microarray dataset, with a number of promising results.

9.1.2 Calculation of Network Level and Node Level Metrics in the Networks Inferred

In previous studies, such as those by Allen et al [94], and Hurley et al [9], the performance of a number of network inference methods for microarray data was compared. However, the validation of these results did not employ a gene ranking system for the particular disease of the microarray dataset. The second objective addresses this; for each of the GRNs inferred in this work, high ranking genes for a number of graph theory metrics are identified, and these genes are scored based on a ranked gene list for the disease returned by a text-mining tool. This is done with the aim of further developing the approach to network inference method comparison, by introducing a quantitative approach to biological validation. To our knowledge, this is the first attempt to quantitatively score network metrics applied to GRNs based on the genes that they identify. Weighted degree centrality is shown to the best metric at identifying scoring genes across all of the networks, followed by degree centrality. Weighted betweenness is the worst performing. Table 9.1 shows the overall ranks of the metrics across all of the GRNs in this work, based on the Génie scoring genes identified by the metrics for their top 20, top100, and top 500 ranking genes.

TABLE 9.1 OVERALL METRIC RANKS IN THE DIFFERENT CATEGORIES OF GRNS ACROSS ALL OF THE MICROARRAY DATASETS

	WGCNA GBM GRNs	Novel Method GBM GRNs	Wang Dataset GRNs	Molenaar Dataset GRNs	Positive Z score RB GRNs	Negative Z score RB GRNs	Ranking totals	Overall rank
Weighted Degree	4th	1st	2nd	1st	1st	3rd	12	1st
Degree Centrality	5th	1st	1st	1st	5th	3rd	16	2nd
Combined Metric	6th	3rd	3rd	4th	2nd	1st	19	3rd
Weighted Closeness	2nd	5th	6th	6th	3rd	2nd	24	4th
Eigenvector Centrality	3rd	6th	5th	1st	4th	6th	25	5th
Weighted Betweenness	1st	4th	4th	5th	6th	5th	25	5th

Furthermore, network level metrics were calculated in all of the networks inferred, and an interesting phenomenon was observed with the value of the diameter in the networks inferred using all variants of the Z score method. The network inferred from the most advanced stage of the disease for the glioblastoma, both neuroblastoma, and the retinoblastoma microarray datasets had the greatest diameter value compared to the networks inferred for the other disease stages in the respective datasets. In section 2.5, low diameter in a GRN was proposed as an indicator of genes being able to communicate with each other easily. One hypothesis is that the high diameter value observed is a contributing factor to the most advanced stage of the disease; certain genes, such as tumour suppressor genes, are not able to easily communicate with other genes involved in regulating tumour progression.

9.1.3 Comparison of GRNs inferred from Two Neuroblastoma Microarray Datasets and Comparison of Disease Stage GRNs to a Healthy GRN inferred from Retinoblastoma Microarray Dataset

As well as the inference and analysis of GRNs for different stages in a microarray dataset, GRNs were also inferred for the same disease categories across different microarray datasets for the same disease, namely neuroblastoma. This was set out as the third objective of the work. Although three out of the four disease stages common to both neuroblastoma datasets did not show any notable correlation, the 4M disease stage did. 7 out of the top 20 ranked genes for the combined metric were common to both GRNs. Previous network inference studies for disease have shown little concordance with other network inference approaches for the same disease microarray dataset, so this represents a promising finding. Another promising result was MYCN as the top ranked gene, based on the combined metric, for the GRNs inferred from the stage 4M samples in both datasets, something specifically commented on by the biologists.

Extending both the first and third objectives of the work, a refined version of the novel network inference method was used to infer GRNs from a retinoblastoma microarray dataset containing a small number of both previously categorised retinoblastoma subtype and normal retinal samples. This extended the work in the thesis, by comparing disease stage GRNs to a normal GRN. A number of significant results were returned, specifically the identification of the gene E2F1 as a highly ranked gene in the normal GRN inferred using the negative Z score approach. Furthermore, the small number of samples in this microarray dataset had previously prevented the use of GRN inference and analysis using existing methods. The use of the novel network inference method made it possible to compare two retinoblastoma subtypes to normal retinal samples using a GRN inference and analysis approach. This potentially allows biologists and clinicians to infer and analyse GRNs for microarray data they had previously been unable to analyse, as a number of microarray datasets contain very few samples per disease stage. This specific example highlights one of the main advantages of the novel network inference method, namely that it does not require a large number of samples to infer a GRN.

9.2 Limitations

Whilst this work has provided a number of contributions, it does have certain limitations. The first limitation concerns the underlying assumption of inferring networks that have a power law degree distribution. Whilst this property has been observed in a number of other networks, including biological networks, it is an assumption of this work that all GRNs display this property. As with any model, it is likely that assumptions used do not always hold true, and that as such GRNs do not always display this property.

The second limitation of the work is that directionality is not defined for the GRNs that are inferred. Directionality has the potential to further identify genes of interest, as genes with a high out-degree are likely to be important genes in the network as this is suggestive of controlling the expression level of a number of other genes. It should however be remembered that a number of other methods for network inference that assign directionality have been shown to be inaccurate, as noted by Hurley et al [9].

The third limitation is the lack of overlap in the top scoring genes between the common disease stage GRNs inferred for neuroblastoma. The best overlap in results is observed for the common stage 4M GRNs, classified using the expression level of MYCN. This is an obvious limitation of the work, as it does suggest that there are some issues with comparing disease stage networks from different datasets using the method presented in this thesis. This could be due to a number of factors. The first is that an additional process might be required in order to normalise microarray datasets from different sources for comparison before networks are inferred. Considering that the two microarray platforms for the neuroblastoma datasets are somewhat different from each other, especially in the number of probes, this is a pertinent issue. The second is the implication that using expression based criteria is better suited to classifying microarray data, rather than using either clinical stage or other evolutionary information. This would again require an additional step to classify the samples in the microarray dataset, and would require the identification and implication of suitable gene signatures for classification. There is also the possibility that this might not be of use to biologists and clinicians, who are more interested in the networks of disease stage, rather than networks based on a different classification criteria.

9.3 Future Work

Whilst a number of contributions to the field have been outlined in this work, there are a number of areas that should be taken into consideration for future work. The first point to note is the availability of samples across all disease stages for certain tumours. For certain tumours, such as retinoblastoma, there is very low availability of low disease stage samples. In the absence of low stage samples, it is not possible to infer GRNs for these low stages. Whilst in this work this has been overcome to an extent by comparing the high stage GRNs inferred to a reference normal GRN, it is something to be aware, and also a restriction meaning that the approach adopted in this work may not be feasible for all tumours due to the absence of low stage samples.

The second point to note concerns the method by which the genes identified by the metrics, and also in the largest cliques in the GRNs, have been scored. The text-mining tool Génie was used as a disease-agnostic approach for quantitatively scoring the genes identified. However, this tool does not provide any evolutionary or disease stage detail to the results. For example; a gene may have been identified in a number of studies as being associated with stage 4 of a disease, but the identification of this gene in a GRN inferred for stage 1 of that disease by a metric will still result in that metric being highly scored. As such, a tool that is able to provide additional disease and evolutionary specific detail to the genes identified would greatly aid in the quantitative scoring of the metrics, and would result in greater biological accuracy for the second objective of this work.

Linked to this is the use of enrichment databases, specifically GeneSetDB. Whilst GeneSetDB has not been used to quantitatively score the genes identified by the metrics, it has been used to interpret results. One possible approach could be to use enrichment databases as a disease-agnostic means by which to score the genes identified, although this will require the subjective interpretation of experts, namely clinicians and biologists. It should also be noted that enrichment databases do not detail all biological processes. In this work, a number of genes either identified as being in the largest cliques or highly ranked did not return any enrichment results. Whilst it was suggested that this could be due to these genes not being involved in any significant biological processes, another reason could be that the biological processes that these genes are involved in are not detailed in the enrichment results. As such, the use of more enrichment databases could counter this issue. For this work, GeneSetDB was chosen as the choice of enrichment database as it offers greater coverage than other enrichment databases; however it could well be that using additional enrichment databases, such as DAVID and GATHER, alongside GeneSetDB would result in additional enrichment results being returned.

During the initial stages of the work, interaction databases, such as BioGRID [147] and IntAct [148], were consulted in an attempt to verify whether interactions in the inferred GRNs were accurate. This is also another potential approach by which to score the genes identified in the networks inferred from the various algorithms. However, this approach was abandoned, as only a small number of interactions have been experimentally validated. As such, even in the case that interactions in inferred networks are completely accurate, there is no way to validate this using the interaction databases. Verification of results is potentially the greatest challenge to bioinformaticians working in the area of network inference from microarray data.

The underlying nature of the microarray dataset that is used to infer the GRNs from should also be taken into account; the data preparation and the platform are greatly influential. In chapter 6, it was noted that there were only 4829 genes in common between the two neuroblastoma microarray datasets. This is a major problem concerning the analysis of microarrays for the same disease across different studies, as newer microarray technology is introduced and utilised, it creates problems for comparisons with older studies that used older microarray platforms. Whilst microarray technology has been available for a number of years, issues of backward compatibility suggest the technology is still immature, and that a flexible protocol is needed. An attempt has been made at incorporating flexibility into the approach used in this work, but even then, a very low number of common genes were identified as being highly ranked across the GRNs for three out of the four disease stages common to both neuroblastoma datasets. Whilst the stage 4M GRNs across both neuroblastoma datasets showed a promising overlap, this categorisation was based on a genomic property, the amplification of the gene MYCN, suggesting that for future work across multiple microarray datasets categorisation based on genomic properties, such as gene amplification or under-expression, might be a better choice of classification criteria than disease stage information already present in the dataset.

Despite one of the objectives of the work specifically concerning the quantitative comparison of the metrics, no single metric was shown to perform better than the others in all of the microarray datasets. Weighted degree and degree centrality, the top two performing metrics, were both the best performing metrics in the GRNs for three out of the six microarray categories, as shown in table 9.1. Additionally, these results do not provide any novel findings; despite adopting what is believed to be a novel approach at identifying appropriate metrics to use for identifying genes of interest across different evolutionary and disease stages of GRNs, the degree centrality is the most commonly studied and used graph theory metric. Perhaps more surprising is that weighted betweenness centrality is the joint worst performing metric overall, considering that it is arguably the second most widely used metric after degree centrality. The finding that the diameter value is greatest in the GRN for the most advanced disease stage in the GRNs inferred from the glioblastoma, both neuroblastoma, and retinoblastoma microarray datasets is believed to be a novel finding, and warrants further investigation.

Finally, in this work it has been shown that adopting a GRN inference and analysis approach for investigating the evolution and progression of tumours has yielded a number of promising results, and has identified specific genes as being associated with specific disease or evolutionary stages. One explicit recommendation for future work is the implementation of the GRN inference and analysis tools used in this work on a personalised publically accessible Galaxy server. Whilst further testing and evaluation of the local host detailed in chapter 8 will be required, providing biologists and clinicians with the means to directly analyse microarray data through the inference and analysis of GRNs for the different stages within the microarray dataset is beneficial, as they are better suited to determine whether the genes that the metrics mark out are biologically of interest. The overall ethos of this work is that the biologists and clinicians are those with the expert knowledge to properly analyse and interpret the results, and as such, should be provided with a tool to provide them with these results.

APPENDIX A

GÉNIE GENE LISTS AND GENESETDB ENRICHMENT RESULTS

The following appendix contains the list of ranked genes returned by the text-mining tool Génie for glioblastoma, neuroblastoma, and retinoblastoma. Additionally, this appendix contains all the enrichment results returned by the enrichment tool GeneSetDB referred to in the main body of the thesis.

TABLE A.1 RANKED GLIOBLASTOMA GENES GENERATED BY GÉNIE

Gene	**Génie Rank**	**Score**	**Gene**	**Génie Rank**	**Score**	**Gene**	**Génie Rank**	**Score**
IDH1	1	299	SMARCB1	48	252	GRIA1	95	205
MGMT	2	298	BCL2	49	251	ANGPT2	96	204
EGFR	3	297	MYC	50	250	BCAN	97	203
IDH2	4	296	NOTCH1	51	249	EGF	98	202
PTEN	5	295	ERCC1	52	248	EZH2	99	201
TP53	6	294	WT1	53	247	SHH	100	200
PROM1	7	293	PMS2	54	246	MAGEC2	101	199
CHI3L1	8	292	MMP9	55	245	CD248	102	198
VEGFA	9	291	NRP1	56	244	KLF6	103	197
ERBB2	10	290	ERCC2	57	243	MET	104	196
OLIG2	11	289	DCX	58	242	TOP2A	105	195
PDGFRA	12	288	BAI1	59	241	RB1	106	194
BIRC5	13	287	LRRC4	60	240	FOXM1	107	193
TERT	14	286	NF1	61	239	NANOG	108	192
CDKN2A	15	285	GSTP1	62	238	MELK	109	191
AKT1	16	284	NDRG2	63	237	TNFSF10	110	190
IL13RA2	17	283	RTEL1	64	236	LGALS3	111	189
TNC	18	282	MIR221	65	235	PTK2	112	188
MKI67	19	281	GLIPR1	66	234	PARP1	113	187
PTGS2	20	280	CXCL12	67	233	YEATS4	114	186
BRAF	21	279	SPARC	68	232	CYR61	115	185
NES	22	278	ATM	69	231	XRCC3	116	184
GLI1	23	277	EPAS1	70	230	KCNMA1	117	183
IGFBP2	24	276	MIF	71	229	NBN	118	182
GFAP	25	275	MSH2	72	228	MMP19	119	181
CA9	26	274	GSTT1	73	227	MSI1	120	180
PLAUR	27	273	PDGFA	74	226	PRKCI	121	179
HIF1A	28	272	TYMS	75	225	MAPK3	122	178
MDM2	29	271	PLAU	76	224	PRND	123	177
SOX2	30	270	MIIP	77	223	AURKA	124	176
AQP4	31	269	CDK4	78	222	TLR9	125	175
CXCR4	32	268	MIR222	79	221	ABCB1	126	174
EPHA2	33	267	SOX10	80	220	CA12	127	173
FABP7	34	266	CDK6	81	219	SOX6	128	172
KDR	35	265	IGF2BP3	82	218	PIK3CG	129	171
SPP1	36	264	AKT2	83	217	ROS1	130	170
PIK3CA	37	263	MTDH	84	216	BAX	131	169
STAT3	38	262	APEX1	85	215	CDKN1B	132	168
CCND1	39	261	NF2	86	214	PTPRM	133	167
KIT	40	260	MMP1	87	213	RAD51	134	166
BMI1	41	259	PHF3	88	212	LGI1	135	165
DMBT1	42	258	SLC7A5	89	211	ASIC1	136	164
CTNNB1	43	257	PHF20	90	210	ABCG2	137	163
IL24	44	256	GSTM1	91	209	TNFRSF12A	138	162
RAC1	45	255	TP73	92	208	DKK1	139	161
PTN	46	254	S100B	93	207	ASPM	140	160
PGR	47	253	NTRK1	94	206	IGF1	141	159

Gene	Génie Rank	Score	Gene	Génie Rank	Score	Gene	Génie Rank	Score
NUMB	142	158	WNT1	191	109	EIF4E	240	60
PTPRZ1	143	157	LRIG3	192	108	LRIG1	241	59
TNFRSF10B	144	156	SOD3	193	107	STAT6	242	58
MIR21	145	155	XRCC1	194	106	SOD2	243	57
RECK	146	154	NQO1	195	105	RHOA	244	56
CTSB	147	153	CD24	196	104	RET	245	55
AURKB	148	152	AXIN1	197	103	LGALS1	246	54
EFNA1	149	151	PHLDB1	198	102	APC	247	53
CTNNBIP1	150	150	BRCA1	199	101	CD34	248	52
DPP4	151	149	GDF15	200	100	PTK2B	249	51
HGF	152	148	ITGB1	201	99	CSF2	250	50
ROBO1	153	147	SLC9A3R1	202	98	PHLPP1	251	49
ERBB3	154	146	L2HGDH	203	97	H2AFX	252	48
SOCS3	155	145	PYGO2	204	96	FHIT	253	47
DKK3	156	144	MIR196A1	205	95	CDH11	254	46
TSC2	157	143	CSF3	206	94	MIA	255	45
XPC	158	142	MAPK1	207	93	MIB1	256	44
CFLAR	159	141	CCR7	208	92	PRKCA	257	43
ING1	160	140	PTCH1	209	91	AQP9	258	42
NFKB1	161	139	BECN1	210	90	CLCN3	259	41
IGF1R	162	138	MIR181A1	211	89	MSH6	260	40
CASP8	163	137	LPO	212	88	MXI1	261	39
RAC2	164	136	PEG3	213	87	BCCIP	262	38
AKT3	165	135	MIR451A	214	86	ID4	263	37
CDKN2B	166	134	IL8	215	85	ITGA7	264	36
TGFB1	167	133	ALKBH2	216	84	MUC6	265	35
FPR1	168	132	ERCC5	217	83	QKI	266	34
RASSF1	169	131	ERBB4	218	82	IMP3	267	33
CD44	170	130	SLC22A18	219	81	VIM	268	32
EZR	171	129	NR2E1	220	80	DCT	269	31
KIAA1549	172	128	TGFB2	221	79	HDGF	270	30
MMP2	173	127	NFIX	222	78	PFKFB3	271	29
MTHFR	174	126	EPOR	223	77	WHSC1	272	28
SLC2A1	175	125	YBX1	224	76	NODAL	273	27
GPR26	176	124	FGF2	225	75	BSG	274	26
GLTSCR1	177	123	BCL2L12	226	74	PDGFRB	275	25
PCDHGA11	178	122	MIR10B	227	73	XRCC5	276	24
HSPA5	179	121	ESR1	228	72	MLH1	277	23
CDKN1A	180	120	CX3CR1	229	71	FLT3LG	278	22
ALDH1A1	181	119	CAV1	230	70	GRPR	279	21
ALAD	182	118	GLTSCR2	231	69	RAD51B	280	20
CDH1	183	117	ALOX5	232	68	CSPG4	281	19
PTENP1	184	116	PAX6	233	67	PDE4A	282	18
BRCA2	185	115	SERPINE1	234	66	TPX2	283	17
YAP1	186	114	RAC3	235	65	NOS1	284	16
PRKDC	187	113	MAML2	236	64	CLU	285	15
LIG4	188	112	ALK	237	63	MMP7	286	14
VCAN	189	111	NGFR	238	62	MTOR	287	13
EMP3	190	110	MYCN	239	61	RECQL4	288	12

Gene	Génie Rank	Score	Gene	Génie Rank	Score	Gene	Génie Rank	Score
MMP14	289	11	MPG	293	7	MTR	297	3
DOCK1	290	10	CTGF	294	6	LAMC1	298	2
GLUL	291	9	SOX4	295	5	ESR2	299	1
PEA15	292	8	ING4	296	4			

TABLE A.2 ENRICHMENT RESULTS LARGEST CLIQUE 200 OR LESS GLIOBLASTOMA CATEGORY INFERRED USING WGCNA

Class	Set Name	Source_DB	Annotated genes	Non-annotated	p.value
GO	synaptic transmission (GO:0007268)	GO_BP	30	359	1.20E-22
Pathway	Transmission across Chemical Synapses	Reactome	21	169	3.10E-19
Pathway	Neuronal System	Reactome	23	260	5.70E-18
Pathway	Glutamate Neurotransmitter Release Cycle	Reactome	9	7	4.20E-16
Pathway	Neurotransmitter Release Cycle	Reactome	11	25	8.90E-16
GO	glutamate secretion (GO:0014047)	GO_BP	9	8	9.00E-16
Pathway	Dopamine Neurotransmitter Release Cycle	Reactome	8	4	2.80E-15
Pathway	Serotonin Neurotransmitter Release Cycle	Reactome	8	4	2.80E-15
GO	neurotransmitter secretion (GO:0007269)	GO_BP	11	38	4.00E-14
GO	cell junction (GO:0030054)	GO_CC	24	461	7.40E-14

TABLE A.3 ENRICHMENT RESULTS LARGEST CLIQUE 201 – 400 GLIOBLASTOMA CATEGORY INFERRED USING WGCNA

Class	Set Name	Source_DB	Annotated genes	Non-annotated	p.value
GO	synaptic transmission (GO:0007268)	GO_BP	30	359	2.40E-20
Pathway	Transmission across Chemical Synapses	Reactome	21	169	1.20E-17
Pathway	Neuronal System	Reactome	24	259	2.30E-17
GO	neurotransmitter secretion (GO:0007269)	GO_BP	11	38	2.70E-13
Pathway	Glutamate Neurotransmitter Release Cycle	Reactome	8	8	2.80E-13
Pathway	Neurotransmitter Release Cycle	Reactome	10	26	3.20E-13
GO	glutamate secretion	GO_BP	8	9	5.30E-13

	(GO:0014047)				
Disease/Phenotype	convulsive seizures	MPO	10	29	7.80E-13
Pathway	Dopamine Neurotransmitter Release Cycle	Reactome	7	5	2.20E-12
Pathway	Serotonin Neurotransmitter Release Cycle	Reactome	7	5	2.20E-12

TABLE A.4 ENRICHMENT RESULTS LARGEST CLIQUE 401 – 600 GLIOBLASTOMA CATEGORY INFERRED USING WGCNA

Class	Set Name	Source_DB	Annotated genes	Non-annotated	p.value
GO	synaptic transmission (GO:0007268)	GO_BP	26	363	1.90E-18
Pathway	Transmission across Chemical Synapses	Reactome	18	172	1.20E-15
GO	synapse (GO:0045202)	GO_CC	20	264	8.50E-15
Pathway	Neuronal System	Reactome	19	264	1.00E-13
GO	cell junction (GO:0030054)	GO_CC	23	462	3.20E-13
Pathway	Dopamine Neurotransmitter Release Cycle	Reactome	7	5	5.20E-13
Pathway	Serotonin Neurotransmitter Release Cycle	Reactome	7	5	5.20E-13
GO	neurotransmitter secretion (GO:0007269)	GO_BP	10	39	1.20E-12
Pathway	Neurotransmitter Release Cycle	Reactome	9	27	2.30E-12
Pathway	Glutamate Neurotransmitter Release Cycle	Reactome	7	9	7.40E-12

TABLE A.5 ENRICHMENT RESULTS LARGEST CLIQUE 601 – 800 GLIOBLASTOMA CATEGORY INFERRED USING WGCNA

Class	Set Name	Source_DB	Annotated genes	Non-annotated	p.value
GO	synaptic transmission (GO:0007268)	GO_BP	37	352	2.00E-26
Pathway	Transmission across Chemical Synapses	Reactome	24	166	3.00E-20
Pathway	Neuronal System	Reactome	27	256	1.90E-19
GO	synapse (GO:0045202)	GO_CC	24	260	3.60E-16
Disease/ Phenotype	abnormal CNS synaptic transmission	MPO	16	86	2.20E-15
GO	neurotransmitter secretion	GO_BP	12	37	2.70E-14

	(GO:0007269)				
Pathway	Glutamate Neurotransmitter Release Cycle	Reactome	8	8	7.10E-13
Pathway	Neurotransmitter Release Cycle	Reactome	10	26	9.90E-13
GO	glutamate secretion (GO:0014047)	GO_BP	8	9	1.30E-12
Pathway	Dopamine Neurotransmitter Release Cycle	Reactome	7	5	4.90E-12

TABLE A.6 ENRICHMENT RESULTS LARGEST CLIQUE MORE THAN 800 GLIOBLASTOMA CATEGORY INFERRED USING WGCNA

Class	Set Name	Source_DB	Annotated genes	Non-annotated	p.value
GO	synaptic transmission (GO:0007268)	GO_BP	21	368	5.30E-20
Pathway	Neuronal System	Reactome	14	269	6.00E-13
Pathway	Transmission across Chemical Synapses	Reactome	12	178	1.80E-12
Drug/ Chemical	gamma-aminobutyric acid(CID000000119)	STITCH	10	199	2.20E-09
GO	neurotransmitter secretion (GO:0007269)	GO_BP	6	43	1.60E-08
Drug/ Chemical	gamma-aminobutyric acid(CID100000119)	STITCH	9	199	3.60E-08
Disease/ Phenotype	abnormal CNS synaptic transmission	MPO	7	95	5.70E-08
Pathway	Neurotransmitter Release Cycle	Reactome	5	31	1.40E-07
GO	synaptic vesicle membrane (GO:0030672)	GO_CC	5	38	3.50E-07
Pathway	Neurotransmitter Receptor Binding And Downstream Transmission In The Postsynaptic Cell	Reactome	7	129	4.10E-07

TABLE A.7 TOP TEN GENESETDB ENRICHMENT RESULTS FOR COMBINED RANK TOP 20 GENES IN LESS THAN 200 GRN CATEGORY INFERRED USING WGCNA

Class	Set Name	Source_DB	Annotated genes	Non-annotated	p.value
Pathway	Neurotransmitter Release Cycle	Reactome	5	31	1.70E-10
Pathway	Dopamine Neurotransmitter Release Cycle	Reactome	4	8	3.00E-10
Pathway	Serotonin Neurotransmitter	Reactome	4	8	3.00E-10

	Release Cycle				
GO	neurotransmitter secretion (GO:0007269)	GO_BP	5	44	8.50E-10
Pathway	Glutamate Neurotransmitter Release Cycle	Reactome	4	12	1.10E-09
GO	glutamate secretion (GO:0014047)	GO_BP	4	13	1.40E-09
Disease/ Phenotype	abnormal neurotransmitter secretion	MPO	4	17	3.60E-09
Pathway	Transmission across Chemical Synapses	Reactome	6	184	1.70E-08
GO	synaptic vesicle membrane (GO:0030672)	GO_CC	4	39	7.30E-08
Pathway	Acetylcholine Neurotransmitter Release Cycle	Reactome	3	8	1.20E-07

TABLE A.8 TOP TEN GENESETDB ENRICHMENT RESULTS FOR COMBINED RANK TOP 20 GENES IN 201-400 SURVIVAL DAYS GRN CATEGORY INFERRED USING WGCNA

Class	**Set Name**	**Source_DB**	**Annotated genes**	**Non-annotated**	**p.value**
Disease/ Phenotype	Syndromic X-linked mental retardation with epilepsy or seizures, including:	KEGG(Disease)	2	18	1.10E-04
GO	cellular response to drug (GO:0035690)	GO_BP	2	23	1.80E-04
GO	cellular response to transforming growth factor beta stimulus (GO:0071560)	GO_BP	2	24	1.90E-04
GO	antiporter activity (GO:0015297)	GO_MF	2	37	4.40E-04
Disease/ Phenotype	Intellectual disability, progressive	HPO	2	54	9.10E-04
Pathway	Platelet homeostasis	Reactome	2	80	1.90E-03
Pathway	Transport of inorganic cations/anions and amino acids/oligopeptides	Reactome	2	93	2.60E-03
Pathway	ErbB1 downstream signaling	PID	2	104	3.20E-03
GO	sodium ion transport (GO:0006814)	GO_BP	2	117	4.00E-03
GO	perinuclear region of cytoplasm (GO:0048471)	GO_CC	3	427	4.50E-03

TABLE A.9 TOP TEN GENESETDB ENRICHMENT RESULTS FOR COMBINED RANK TOP 20 GENES IN 401-600 SURVIVAL DAYS GRN CATEGORY INFERRED USING WGCNA

Class	Set Name	Source_DB	Annotated genes	Non-annotated	p. value
Pathway	Neurotransmitter Release Cycle	Reactome	3	33	7.40E-06
GO	neurotransmitter secretion (GO:0007269)	GO_BP	3	46	1.90E-05
Drug/Chemical	aminooxyacetic acid(CID100000285)	STITCH	2	8	4.90E-05
Disease/ Phenotype	decreased aggression towards males	MPO	2	12	9.90E-05
Drug/Chemical	aminooxyacetic acid(CID000000285)	STITCH	2	12	9.90E-05
GeneRegulation	ATF4	TFactS	2	16	1.70E-04
GO	synapse (GO:0045202)	GO_CC	4	280	2.10E-04
Pathway	Glutamate_Glutamine_ metabolism	INOH	2	24	3.50E-04
GO	synaptic transmission (GO:0007268)	GO_BP	4	385	6.90E-04
Disease/ Phenotype	impaired glucose tolerance	MPO	3	180	9.40E-04

TABLE A.10 TOP TEN GENESETDB ENRICHMENT RESULTS FOR COMBINED RANK TOP 20 GENES IN 601-800 SURVIVAL DAYS GRN CATEGORY INFERRED USING WGCNA

Class	Set Name	Source_DB	Annotated genes	Non-annotated	p.value
GO	calcium ion-dependent exocytosis (GO:0017156)	GO_BP	2	10	7.20E-05
GO	regulation of synaptic plasticity (GO:0048167)	GO_BP	2	18	2.10E-04
GO	cellular response to drug (GO:0035690)	GO_BP	2	23	3.20E-04
Pathway	Insulin-mediated glucose transport	PID	2	27	4.40E-04
Pathway	Calcium Regulation in the Cardiac Cell(WP536)	WikiPathways	3	148	5.40E-04
GO	calcium ion transmembrane transport (GO:0070588)	GO_BP	2	32	6.00E-04
GO	calmodulin binding (GO:0005516)	GO_MF	3	156	6.20E-04
GO	synaptic vesicle membrane (GO:0030672)	GO_CC	2	41	9.60E-04
GO	hydrolase activity, acting on acid anhydrides, catalyzing transmembrane movement of substances (GO:0016820)	GO_MF	2	43	1.10E-03
GO	protein complex (GO:0043234)	GO_CC	3	193	1.10E-03

TABLE A.11 TOP TEN GENESETDB ENRICHMENT RESULTS FOR COMBINED RANK TOP 20 GENES IN MORE THAN 800 SURVIVAL DAYS GRN CATEGORY INFERRED USING WGCNA

Class	Set Name	Source_DB	Annotated genes	Non-annotated	p.value
GO	synaptic transmission (GO:0007268)	GO_BP	9	380	9.30E-11
Pathway	Transmission across Chemical Synapses	Reactome	6	184	3.50E-08
Pathway	Neuronal System	Reactome	6	277	3.70E-07
Drug/ Chemical	gamma-aminobutyric acid(CID100000119)	STITCH	5	203	2.20E-06
Drug/ Chemical	gamma-aminobutyric acid(CID000000119)	STITCH	5	204	2.30E-06
Pathway	Neurotransmitter Release Cycle	Reactome	3	33	7.40E-06
GO	cell junction (GO:0030054)	GO_CC	6	479	8.50E-06
Disease/ Phenotype	convulsive seizures	MPO	3	36	9.40E-06
GO	synapse (GO:0045202)	GO_CC	5	279	1.00E-05
Drug/ Chemical	muscimol(CID100004266)	STITCH	3	40	1.30E-05

TABLE A.12 ENRICHMENT RESULTS LARGEST CLIQUE 200 OR LESS GLIOBLASTOMA CATEGORY INFERRED USING THE NOVEL Z-SCORE INFERENCE METHOD

Class	Set Name	Source_DB	Annotated genes	Non-annotated	p.value
GO	extracellular matrix (GO:0031012)	GO_CC	21	132	4.00E-25
Pathway	Beta1 integrin cell surface interactions	PID	10	55	5.90E-13
GO	extracellular matrix structural constituent (GO:0005201)	GO_MF	10	56	6.90E-13
Disease/ Phenotype	abnormal cutaneous collagen fibril morphology	MPO	7	13	3.50E-12
Pathway	Syndecan-1-mediated signaling events	PID	8	38	4.80E-11
Gene Regulation	SMAD7	TFactS	6	9	5.10E-11
Disease/ Phenotype	Dermal atrophy	HPO	7	30	4.40E-10
Disease/ Phenotype	Mitral valve prolapse	HPO	6	16	7.50E-10
Disease/ Phenotype	Osteoarthritis	HPO	6	16	7.50E-10
Disease/ Phenotype	abnormal tendon morphology	MPO	6	17	1.00E-09

TABLE A.13 ENRICHMENT RESULTS LARGEST CLIQUE 201-400 GLIOBLASTOMA CATEGORY INFERRED USING THE NOVEL Z-SCORE INFERENCE METHOD

Class	**Set Name**	**Source_DB**	**Annotated genes**	**Non-annotated**	**p.value**
Pathway	EPHB forward signalling	PID	6	30	1.90E-07
Pathway	Internalization of ErbB1	PID	6	34	3.70E-07
Drug/Chemical	inositol 1,4,5-trisphosphate (CID000439456)	STITCH	9	121	3.80E-07
Pathway	CXCR3-mediated signaling events	PID	6	36	5.00E-07
Drug/Chemical	inositol 1,4,5-trisphosphate (CID100000806)	STITCH	9	130	6.80E-07
Pathway	Dopamine Neurotransmitter Release Cycle	Reactome	4	8	1.20E-06
Pathway	Serotonin Neurotransmitter Release Cycle	Reactome	4	8	1.20E-06
GO	neurotransmitter secretion (GO:0007269)	GO_BP	6	43	1.30E-06
Pathway	Role of ?-arrestins in the activation and targeting of MAP kinases	Biocarta	4	10	2.30E-06
Pathway	Purine metabolism	EHMN	11	255	3.30E-06

TABLE A.14 ENRICHMENT RESULTS LARGEST CLIQUE 401-600 GLIOBLASTOMA CATEGORY INFERRED USING THE NOVEL Z-SCORE INFERENCE METHOD

Class	**Set Name**	**Source_DB**	**Annotated genes**	**Non-annotated**	**p.value**
Disease/ Phenotype	increased susceptibility to bacterial infection	MPO	15	193	4.10E-12
GO	interferon-gamma-mediated signaling pathway (GO:0060333)	GO_BP	10	59	1.70E-11
Disease/ Phenotype	abnormal macrophage physiology	MPO	14	191	4.60E-11
Gene Regulation	SPI1	TFactS	10	77	1.80E-10
Disease/ Phenotype	abnormal antigen presenting cell physiology	MPO	6	10	4.30E-10
Disease/ Phenotype	increased IgG level	MPO	9	62	6.40E-10
Pathway	Interferon gamma signaling	Reactome	9	64	8.20E-10
Drug/Chemical	N-acetylglucosamine (CID100000899)	STITCH	16	363	2.20E-09
Disease/ Phenotype	decreased susceptibility to bacterial infection	MPO	8	61	1.20E-08
Disease/ Phenotype	increased lymphocyte cell number	MPO	8	62	1.40E-08

TABLE A.15 ENRICHMENT RESULTS LARGEST CLIQUE 601-800 GLIOBLASTOMA CATEGORY INFERRED USING THE NOVEL Z-SCORE INFERENCE METHOD

Class	Set Name	Source_DB	Annotated genes	Non-annotated	p.value
GO	synaptic transmission (GO:0007268)	GO_BP	32	357	1.20E-18
Pathway	Transmission across Chemical Synapses	Reactome	22	168	3.20E-16
Pathway	Neuronal System	Reactome	23	260	1.50E-13
GO	synapse (GO:0045202)	GO_CC	22	262	1.40E-12
GO	cell junction (GO:0030054)	GO_CC	28	457	1.40E-12
GO	postsynaptic membrane (GO:0045211)	GO_CC	17	152	8.60E-12
Drug/Chemical	glutamate(CID100000611)	STITCH	21	287	5.10E-11
GO	postsynaptic density (GO:0014069)	GO_CC	13	87	1.00E-10
Drug/Chemical	glutamate(CID000033032)	STITCH	21	306	1.50E-10
Pathway	Neurotransmitter Receptor Binding And Downstream Transmission In The Postsynaptic Cell	Reactome	14	122	4.60E-10

TABLE A.16 ENRICHMENT RESULTS LARGEST CLIQUE MORE THAN 800 GLIOBLASTOMA CATEGORY INFERRED USING THE NOVEL Z-SCORE INFERENCE METHOD

Class	Set Name	Source_DB	Annotated genes	Non-annotated	p.value
Disease/ Phenotype	anophthalmia	MPO	6	58	3.10E-06
Drug/Chemical	MT19c compound	CTD	13	486	1.70E-05
Drug/Chemical	N-acetylsphingosine	CTD	4	22	2.00E-05

TABLE A.17 ENRICHMENT RESULTS COMBINED RANK TOP 20 GENES IN LESS THAN 200 GRN CATEGORY INFERRED USING THE NOVEL Z-SCORE INFERENCE METHOD

Class	Set Name	Source_DB	Annotated genes	Non-annotated	p.value
Disease/ Phenotype	partial neonatal lethality	MPO	5	310	1.70E-05

TABLE A.18 ENRICHMENT RESULTS COMBINED RANK TOP 20 GENES IN 201 -400 GRN CATEGORY INFERRED USING THE NOVEL Z-SCORE INFERENCE METHOD

Class	Set Name	Source_DB	Annotated genes	Non-annotated	p.value
Pathway	GMCSF-mediated signaling events	PID	3	32	6.70E-06
Pathway	Angiopoietin receptor Tie2-mediated signaling	PID	3	45	1.80E-05
Pathway	IL-4 signaling pathway(WP395)	WikiPathways	3	48	2.10E-05
Pathway	Fc-epsilon receptor I signaling in mast cells	PID	3	55	3.10E-05
Pathway	Leptin signaling pathway(WP2034)	WikiPathways	3	58	3.60E-05

Pathway	BCR signaling pathway	PID	3	63	4.60E-05
Drug/Chemical	di-arsenic-trioxide(CID100518740)	STITCH	2	9	6.00E-05
Drug/Chemical	di-arsenic-trioxide(CID000518740)	STITCH	2	9	6.00E-05
Pathway	Osteopontin Signaling(WP1434)	WikiPathways	2	9	6.00E-05
Pathway	IL4 Signaling Pathway	NetPath	3	72	6.80E-05

TABLE A.19 ENRICHMENT RESULTS COMBINED RANK TOP 20 GENES IN 401 -600 GRN CATEGORY INFERRED USING THE NOVEL Z-SCORE INFERENCE METHOD

Class	**Set Name**	**Source_DB**	**Annotated genes**	**Non-annotated**	**p.value**
Disease/ Phenotype	rib bifurcation	MPO	2	26	4.10E-04
Disease/ Phenotype	squamous cell carcinoma	MPO	2	33	6.40E-04
GeneRegulation	TP53	TFactS	3	142	4.80E-04
GO	response to morphine (GO:0043278)	GO_BP	2	18	2.10E-04
GO	response to UV (GO:0009411)	GO_BP	2	34	6.70E-04
Pathway	DNA damage response(WP707)	WikiPathways	3	66	5.30E-05
Pathway	Direct p53 effectors	PID	3	133	3.90E-04

TABLE A.20 ENRICHMENT RESULTS COMBINED RANK TOP 20 GENES IN MORE THAN 800 GRN CATEGORY INFERRED USING THE NOVEL Z-SCORE INFERENCE METHOD

Class	**Set Name**	**Source_DB**	**Annotated genes**	**Non-annotated**	**p.value**
Pathway	Benzo(a)pyrene metabolism(WP696)	WikiPathways	2	8	Pathway
GO	tight junction (GO:0005923)	GO_CC	3	88	GO
Disease/ Phenotype	prolonged diestrus	MPO	2	17	Disease/Phenotype
Disease/ Phenotype	prolonged estrous cycle	MPO	2	29	Disease/Phenotype
Disease/ Phenotype	disorganized embryonic tissue	MPO	2	31	Disease/Phenotype
Pathway	Regulation of RAC1 activity	PID	2	36	Pathway
Pathway	E-cadherin signaling in the nascent adherens junction	PID	2	36	Pathway
GO	cytoskeletal protein binding (GO:0008092)	GO_MF	2	42	GO
GO	cell cycle (GO:0007049)	GO_BP	4	430	GO
GO	negative regulation of neuron differentiation (GO:0045665)	GO_BP	2	45	GO

Table A.21 Ranked Neuroblastoma Genes Generated by Génie

Gene	Génie Rank	Score	Gene	Génie Rank	Score	Gene	Génie Rank	Score
MYCN	1	500	NME1	48	453	RET	95	406
EGFR	2	499	IGF1	49	452	GSK3B	96	405
AKT1	3	498	ID1	50	451	ADAM10	97	404
VEGFA	4	497	BCL2	51	450	TP73	98	403
PTGS2	5	496	CCND1	52	449	SHH	99	402
TP53	6	495	ITGB1	53	448	UBC	100	401
TGFB1	7	494	IL8	54	447	ABCB1	101	400
MMP2	8	493	PTN	55	446	TGFA	102	399
APP	9	492	DKK1	56	445	FAS	103	398
MMP9	10	491	MMP1	57	444	MDM2	104	397
ALK	11	490	IL24	58	443	HYAL1	105	396
CD44	12	489	BSG	59	442	RUNX2	106	395
MMP14	13	488	PLAUR	60	441	SPARC	107	394
NFKB1	14	487	CDH1	61	440	RB1	108	393
TNF	15	486	CD4	62	439	MDK	109	392
CXCR4	16	485	IL6	63	438	BAX	110	391
STAT3	17	484	PIK3CG	64	437	AR	111	390
NTRK1	18	483	CTGF	65	436	MMP7	112	389
HIF1A	19	482	PTHLH	66	435	MAPK8	113	388
BMP2	20	481	L1CAM	67	434	SOD1	114	387
PLAU	21	480	KDR	68	433	CCL2	115	386
ERBB2	22	479	SP1	69	432	IGFBP7	116	385
HGF	23	478	TGM2	70	431	CTSB	117	384
TERT	24	477	ITGAV	71	430	BMP7	118	383
MAPK1	25	476	PTK2	72	429	RAC1	119	382
IGF1R	26	475	ID2	73	428	PIK3CA	120	381
HPSE	27	474	BIRC5	74	427	CDKN1B	121	380
PTEN	28	473	CDKN2A	75	426	ALOX12	122	379
MYC	29	472	VEGFC	76	425	PRNP	123	378
FGF2	30	471	SNCA	77	424	TNFRSF11B	124	377
TNFSF10	31	470	CASP8	78	423	HMOX1	125	376
MAPK3	32	469	SERPINB5	79	422	PSEN1	126	375
NOTCH1	33	468	PPARG	80	421	THBS1	127	374
CDKN1A	34	467	PDGFB	81	420	PRKCD	128	373
PROM1	35	466	BMP4	82	419	MCAM	129	372
ADAM17	36	465	MAPT	83	418	NANOG	130	371
MTOR	37	464	BRAF	84	417	PRKCA	131	370
NCAM1	38	463	KIT	85	416	NF2	132	369
CAV1	39	462	SPP1	86	415	FASN	133	368
SRC	40	461	CSF1	87	414	NT5E	134	367
TNFSF11	41	460	INHBA	88	413	ALOX5	135	366
NTRK2	42	459	PHOX2B	89	412	AREG	136	365
CXCL12	43	458	TIMP1	90	411	DKK3	137	364
MET	44	457	CADM1	91	410	GDNF	138	363
FN1	45	456	EGR1	92	409	KITLG	139	362
CTNNB1	46	455	MAPK14	93	408	CYR61	140	361
MUC1	47	454	HBEGF	94	407	SERPINE1	141	360

Gene	Génie Rank	Score	Gene	Génie Rank	Score	Gene	Génie Rank	Score
TNFRSF10B	142	359	ADM	191	310	ABCG2	240	261
F3	143	358	FGFR2	192	309	MIR34A	241	260
SLC7A5	144	357	BDNF	193	308	SHC1	242	259
EDN1	145	356	IFNG	194	307	NOS1	243	258
CASP3	146	355	SLC3A2	195	306	FYN	244	257
TH	147	354	CTSD	196	305	HRAS	245	256
FGFR1	148	353	SMAD3	197	304	ICAM1	246	255
NF1	149	352	TWIST1	198	303	CDH2	247	254
ESR1	150	351	PLA2G4A	199	302	LOX	248	253
EGF	151	350	EPAS1	200	301	KRAS	249	252
SDC1	152	349	GPC3	201	300	IDH1	250	251
CA9	153	348	CXCL1	202	299	LIF	251	250
TNC	154	347	RELA	203	298	ILK	252	249
ITGA5	155	346	FIGF	204	297	ANXA2	253	248
CD40	156	345	DLK1	205	296	NOV	254	247
SMAD4	157	344	IL1B	206	295	EIF2AK2	255	246
POU5F1	158	343	LAMC2	207	294	RARA	256	245
NGFR	159	342	TFPI2	208	293	THY1	257	244
CYP19A1	160	341	PDGFRB	209	292	NES	258	243
PARK2	161	340	LAMA1	210	291	CNTN1	259	242
RHOA	162	339	HLA-G	211	290	GDF15	260	241
CD24	163	338	CEACAM5	212	289	XIAP	261	240
S100A4	164	337	ITGB3	213	288	NID1	262	239
BMI1	165	336	RASSF1	214	287	PDPN	263	238
CD9	166	335	ERBB4	215	286	KLF4	264	237
FLT1	167	334	EZR	216	285	IGF2	265	236
MFI2	168	333	JUN	217	284	EWSR1	266	235
PTK2B	169	332	SEMA3F	218	283	NDRG1	267	234
NEU3	170	331	ALOX15	219	282	AGER	268	233
EPCAM	171	330	CSF2	220	281	LAMA3	269	232
GJA1	172	329	NRG1	221	280	CD40LG	270	231
PTP4A3	173	328	BACE1	222	279	KLK3	271	230
FOLH1	174	327	PDGFRA	223	278	COL4A2	272	229
IDO1	175	326	MCL1	224	277	PRKCZ	273	228
HDAC1	176	325	TFAP2A	225	276	EPHB2	274	227
MIF	177	324	VIM	226	275	CSPG4	275	226
HSPG2	178	323	SPINT1	227	274	E2F1	276	225
ID3	179	322	PDGFA	228	273	LAMA5	277	224
SERPINF1	180	321	BCL2L1	229	272	FOXO3	278	223
MME	181	320	TIMP2	230	271	IL10	279	222
COL4A1	182	319	ENPP2	231	270	WT1	280	221
HSP90AA1	183	318	COL1A1	232	269	MTAP	281	220
DPP4	184	317	CD82	233	268	ST8SIA1	282	219
CEACAM1	185	316	CDK2	234	267	BCL2L11	283	218
JAK2	186	315	IL2	235	266	MGMT	284	217
LGALS3	187	314	CSF1R	236	265	IGFBP5	285	216
ERBB3	188	313	SP3	237	264	PTPRZ1	286	215
CLU	189	312	CD34	238	263	NOS2	287	214
GLI1	190	311	SLIT2	239	262	HLA-A	288	213

Gene	Génie Rank	Score	Gene	Génie Rank	Score	Gene	Génie Rank	Score
STAT5B	289	212	FLT4	338	163	ANGPT1	387	114
ANPEP	290	211	TNFSF12	339	162	DNMT1	388	113
RECK	291	210	PTPN11	340	161	DPYSL3	389	112
F2R	292	209	ACHE	341	160	COL4A6	390	111
IGFBP2	293	208	GRB2	342	159	GAPDH	391	110
HMGB1	294	207	KLRK1	343	158	TNFRSF11A	392	109
EPO	295	206	HPN	344	157	PTTG1	393	108
DSG3	296	205	KRT8	345	156	ITGB5	394	107
COL4A3	297	204	ALDH1A1	346	155	IL15	395	106
PTGS1	298	203	IL13RA2	347	154	ACTB	396	105
NBL1	299	202	GNRHR	348	153	CX3CL1	397	104
TOP2A	300	201	NTRK3	349	152	NRP1	398	103
ODC1	301	200	STAR	350	151	FAP	399	102
MIA	302	199	CD38	351	150	HAS3	400	101
PLD2	303	198	CPM	352	149	LGALS3BP	401	100
PLCG1	304	197	FASLG	353	148	CEBPB	402	99
PARK7	305	196	PCNA	354	147	STAT1	403	98
VIP	306	195	REST	355	146	RHOC	404	97
FOXM1	307	194	JAG1	356	145	WNT5A	405	96
PLD1	308	193	POLA1	357	144	LGALS1	406	95
EBAG9	309	192	IGFBP3	358	143	CSF3	407	94
NFKBIA	310	191	CTAG1B	359	142	CTSL1	408	93
RARB	311	190	SFRP1	360	141	TNFSF13B	409	92
HAS2	312	189	RTN4	361	140	CDK4	410	91
OPRM1	313	188	NEWENTRY	362	139	LCN2	411	90
CDX2	314	187	LRRC4	363	138	DGAT1	412	89
FHIT	315	186	VCAN	364	137	PTGER2	413	88
STMN1	316	185	FABP7	365	136	CRABP2	414	87
KIF1B	317	184	WWOX	366	135	CD63	415	86
HSPA5	318	183	NDN	367	134	OGFR	416	85
FGFR3	319	182	MSLN	368	133	PTPRJ	417	84
WNT1	320	181	CASZ1	369	132	ITGA3	418	83
HPGD	321	180	AKR1B1	370	131	PRKACA	419	82
SERPINA5	322	179	FOLR1	371	130	LAMC1	420	81
ST14	323	178	TP63	372	129	SOX9	421	80
PRKCE	324	177	CFLAR	373	128	COL1A2	422	79
TYMP	325	176	GPI	374	127	FURIN	423	78
MAP2K1	326	175	ING1	375	126	F2RL1	424	77
PIK3R1	327	174	SCD	376	125	TGFBI	425	76
PHB	328	173	HYAL2	377	124	TNFRSF1B	426	75
GAS1	329	172	SOD2	378	123	PTGES	427	74
ADCYAP1	330	171	PODXL	379	122	GCNT1	428	73
RUNX3	331	170	CASP9	380	121	FGF7	429	72
SLC9A1	332	169	TXN	381	120	PEBP1	430	71
HSPB1	333	168	DLC1	382	119	CXCR7	431	70
PTP4A2	334	167	PFKFB3	383	118	ANGPT2	432	69
FGFR4	335	166	NFE2L2	384	117	FGF1	433	68
PLK1	336	165	MMP13	385	116	ABCC1	434	67
CEACAM6	337	164	PAPPA	386	115	ITGA6	435	66

Gene	**Génie Rank**	**Score**	**Gene**	**Génie Rank**	**Score**	**Gene**	**Génie Rank**	**Score**
CHKA	436	65	SKP2	458	43	HES1	480	21
PLA2G10	437	64	ALCAM	459	42	PRKCI	481	20
TLR4	438	63	NTF3	460	41	LDHA	482	19
DPYSL2	439	62	SIRT1	461	40	LEF1	483	18
SERPINE2	440	61	CAV2	462	39	ASPH	484	17
CAMK2G	441	60	PTPRB	463	38	PTGER1	485	16
CHRM3	442	59	NGF	464	37	SDC3	486	15
NRP2	443	58	KRT19	465	36	RPSA	487	14
AKT2	444	57	ELAVL4	466	35	TGFB3	488	13
LAMB1	445	56	NBAS	467	34	SNAI2	489	12
MTDH	446	55	KISS1	468	33	DCN	490	11
ZEB1	447	54	SPHK1	469	32	HDAC6	491	10
INSR	448	53	ABL1	470	31	HOXB7	492	9
ACTN1	449	52	CHAT	471	30	IL6R	493	8
CREB1	450	51	IGFBP1	472	29	TGFBR1	494	7
CEBPA	451	50	CDCP1	473	28	FHL2	495	6
EP300	452	49	ANG	474	27	PTPRD	496	5
RARG	453	48	DDIT3	475	26	DCD	497	4
NDRG2	454	47	GRN	476	25	IL17A	498	3
IL3	455	46	CXCR2	477	24	TM4SF1	499	2
SSTR2	456	45	ROCK1	478	23	CHD5	500	1
EPHB4	457	44	FUT4	479	22			

TABLE A.22 ENRICHMENT RESULTS LARGEST UNIQUE CLIQUE STAGE 1 GRN MOLENAAR DATASET INFERRED USING THE NOVEL Z-SCORE INFERENCE METHOD

Class	**Set Name**	**Source_DB**	**Annotated genes**	**Non-annotated**	**p.value**
Disease/ Phenotype	abnormal humoral immune response	MPO	13	97	2.20E-15
Disease/ Phenotype	abnormal macrophage physiology	MPO	14	191	3.90E-13
Disease/ Phenotype	decreased CD8-positive T cell number	MPO	13	155	5.80E-13
Disease/ Phenotype	decreased CD4-positive T cell number	MPO	13	165	1.20E-12
Disease/ Phenotype	decreased B cell number	MPO	14	215	1.80E-12
Pathway	Cytokine Signaling in Immune system	Reactome	15	270	2.30E-12
Disease/ Phenotype	decreased T cell proliferation	MPO	13	175	2.40E-12
Pathway	TCR signalling	Reactome	9	58	1.70E-11
Disease/ Phenotype	abnormal T cell physiology	MPO	11	122	1.90E-11
GO	cytokine-mediated signaling pathway (GO:0019221)	GO_BP	13	210	2.10E-11

TABLE A.23 ENRICHMENT RESULTS LARGEST UNIQUE CLIQUE STAGE 3 GRN MOLENAAR DATASET INFERRED USING THE NOVEL Z-SCORE INFERENCE METHOD

Class	**Set Name**	**Source_DB**	**Annotated genes**	**Non-annotated**	**p.value**
Drug/ Chemical	methylselenic acid	CTD	19	383	7.20E-11
Drug/ Chemical	Etoposide	CTD	17	328	4.10E-10
Drug/ Chemical	Plant Extracts	CTD	12	204	4.80E-08
Drug/ Chemical	Thiophenes	CTD	6	35	4.30E-07
Disease/ Phenotype	Arthrogryposis multiplex congenita	HPO	5	20	8.10E-07
Drug/ Chemical	Mitoxantrone	CTD	10	189	1.60E-06
Disease/ Phenotype	Congenital contractures	HPO	5	24	1.80E-06
Disease/ Phenotype	Abnormality of the esophagus	HPO	8	114	2.60E-06
Disease/ Phenotype	abnormal cornea morphology	MPO	6	53	3.90E-06
Pathway	Interferon alpha/beta signaling	Reactome	6	58	6.30E-06

TABLE A.24 ENRICHMENT RESULTS LARGEST UNIQUE CLIQUE STAGE 4 GRN MOLENAAR DATASET INFERRED USING THE NOVEL Z-SCORE INFERENCE METHOD

Class	**Set Name**	**Source_DB**	**Annotated genes**	**Non-annotated**	**p.value**
GO	extracellular matrix (GO:0031012)	GO_CC	20	133	3.10E-22
Pathway	Beta1 integrin cell surface interactions	PID	13	52	2.60E-17
Pathway	Syndecan-1-mediated signaling events	PID	10	36	5.80E-14
Disease/ Phenotype	Abnormality of connective tissue	HPO	16	259	2.30E-12
GO	basement membrane (GO:0005604)	GO_CC	10	58	3.80E-12
Disease/ Phenotype	Joint hypermobility	HPO	10	66	1.20E-11
Disease/ Phenotype	Abnormality of the hip	HPO	11	101	2.80E-11
Disease/ Phenotype	Joint dislocation	HPO	9	50	3.30E-11
Disease/ Phenotype	Abnormality of the joints	HPO	17	382	6.30E-11
Disease/ Phenotype	Herniae	HPO	10	80	6.70E-11

TABLE A.25 ENRICHMENT RESULTS LARGEST UNIQUE CLIQUE STAGE 4 GRN MOLENAAR DATASET INFERRED USING THE NOVEL Z-SCORE INFERENCE METHOD

Class	**Set Name**	**Source_DB**	**Annotated genes**	**Non-annotated**	**p.value**
GO	actin binding (GO:0003779)	GO_MF	10	273	1.30E-07
GO	stress fiber (GO:0001725)	GO_CC	5	40	8.10E-07
Pathway	Smooth Muscle Contraction	Reactome	4	18	1.40E-06
Drug/ Chemical	paricalcitol	CTD	5	49	2.00E-06
GO	extracellular matrix (GO:0031012)	GO_CC	7	146	2.00E-06

Drug/ Chemical	methylselenic acid	CTD	10	392	3.10E-06
GO	actin cytoskeleton (GO:0015629)	GO_CC	7	169	5.20E-06
GO	focal adhesion (GO:0005925)	GO_CC	6	114	6.90E-06
Drug/ Chemical	Phosphorus	CTD	5	66	7.90E-06
GO	negative regulation of transforming growth factor beta receptor signaling pathway (GO:0030512)	GO_BP	4	35	1.50E-05

TABLE A.26 ENRICHMENT RESULTS LARGEST UNIQUE CLIQUE STAGE 1 GRN WANG DATASET INFERRED USING THE NOVEL Z-SCORE INFERENCE METHOD

Class	Set Name	Source_DB	Annotated genes	Non-annotated	p.value
GO	RNA processing (GO:0006396)	GO_BP	5	78	7.50E-06

TABLE A.27 ENRICHMENT RESULTS LARGEST UNIQUE CLIQUE STAGE 3 GRN WANG DATASET INFERRED USING THE NOVEL Z-SCORE INFERENCE METHOD

Class	Set Name	Source_DB	Annotated genes	Non-annotated	p.value
Disease/ Phenotype	abnormal macrophage physiology	MPO	14	191	7.20E-11
Gene Regulation	SP1	TFactS	18	400	2.50E-10
Drug/Chemical	Dexamethasone	CTD	16	321	6.60E-10
Drug/Chemical	Lipopolysaccharides	CTD	13	191	8.40E-10
Drug/Chemical	Paclitaxel	CTD	18	458	2.00E-09
Disease/ Phenotype	necrosis	MPO	8	49	3.40E-09
GO	blood coagulation (GO:0007596)	GO_BP	17	439	7.20E-09
Drug/Chemical	rosiglitazone	CTD	13	231	7.40E-09
Pathway	Beta1 integrin cell surface interactions	PID	8	57	9.80E-09
GO	platelet activation (GO:0030168)	GO_BP	12	193	1.00E-08

TABLE A.28 ENRICHMENT RESULTS LARGEST UNIQUE CLIQUE STAGE 4M GRN WANG DATASET INFERRED USING THE NOVEL Z-SCORE INFERENCE METHOD

Class	Set Name	Source_DB	Annotated genes	Non-annotated	p.value
Pathway	Eukaryotic Translation Elongation	Reactome	43	44	7.90E-75
Pathway	Nonsense Mediated Decay Independent of the Exon Junction Complex	Reactome	43	46	3.00E-74
GO	viral transcription (GO:0019083)	GO_BP	42	40	6.50E-74
GO	viral infectious cycle (GO:0019058)	GO_BP	43	48	1.10E-73
Pathway	Eukaryotic Translation	Reactome	42	42	2.60E-73

	Termination				
Pathway	Peptide chain elongation	Reactome	42	42	2.60E-73
Pathway	Viral mRNA Translation	Reactome	42	43	5.20E-73
GO	translational elongation (GO:0006414)	GO_BP	43	51	6.80E-73
GO	translational termination (GO:0006415)	GO_BP	42	44	1.00E-72
Pathway	Influenza Viral RNA Transcription and Replication	Reactome	43	58	3.80E-71

TABLE A.29 ENRICHMENT RESULTS FOR COMBINED RANK TOP 20 GENES IN STAGE 1 GRN INFERRED FROM MOLENAAR DATASET USING THE NOVEL Z-SCORE INFERENCE METHOD

Class	Set Name	Source_DB	Annotated genes	Non-annotated	p.value
Drug/Chemical	Mercury	CTD	5	297	1.40E-05
Drug/Chemical	Choline	CTD	3	42	1.50E-05
Drug/Chemical	Vitamin A	CTD	3	66	5.30E-05
Disease/Phenotype	abnormal joint mobility	MPO	2	9	6.00E-05
GeneRegulation	SP1	TFactS	5	413	6.50E-05
Drug/Chemical	3-(2-hydroxy-4-(1,1-dimethylheptyl)phenyl)-4-(3-hydroxypropyl)cyclohexanol	CTD	2	10	7.20E-05
Pathway	CD40L Signaling Pathway	Biocarta	2	12	9.90E-05
Pathway	TNFR2 Signaling Pathway	Biocarta	2	15	1.50E-04
Drug/Chemical	Cycloheximide	CTD	4	256	1.50E-04
Disease/Phenotype	decreased susceptibility to induced colitis	MPO	2	18	2.10E-04

TABLE A.30 ENRICHMENT RESULTS FOR COMBINED RANK TOP 20 GENES IN STAGE 4 GRN INFERRED FROM MOLENAAR DATASET USING THE NOVEL Z-SCORE INFERENCE METHOD

Class	Set Name	Source_ DB	Annotated genes	Non-annotated	p.value
Disease/ Phenotype	hypertension	MPO	3	39	1.20E-05
Disease/ Phenotype	abnormal adrenaline level	MPO	2	8	4.90E-05

TABLE A.31 ENRICHMENT RESULTS FOR COMBINED RANK TOP 20 GENES IN STAGE 4M GRN INFERRED FROM MOLENAAR DATASET USING THE NOVEL Z-SCORE INFERENCE METHOD

Class	Set Name	Source_ DB	Annotated genes	Non-annotated	p. value
Disease/ Phenotype	Malaise	SIDER	6	458	6.60E-06
Disease/ Phenotype	Rigors	SIDER	5	413	6.50E-05
Drug / Chemical	7MP (CID110062694)	STITCH	2	11	8.50E-05
Drug/ Chemical	7MP (CID010062694)	STITCH	2	11	8.50E-05
GO	chromatin DNA binding (GO:0031490)	GO_MF	2	13	1.10E-04
Disease/ Phenotype	Mental disorder	SIDER	4	258	1.50E-04
Disease/ Phenotype	Hemorrhage	MPO	4	260	1.60E-04
GO	embryonic cranial skeleton morphogenesis (GO:0048701)	GO_BP	2	18	2.10E-04
Disease/ Phenotype	Vogelstein and Kinzler 2004	Cancer Genes	3	109	2.20E-04
Disease/ Phenotype	Hypoplasia of the toes	HPO	2	19	2.30E-04

TABLE A.32 ENRICHMENT RESULTS FOR COMBINED RANK TOP 20 GENES IN STAGE 1 GRN INFERRED FROM WANG DATASET USING THE NOVEL Z-SCORE INFERENCE METHOD

Class	Set Name	Source_DB	Annotated genes	Non-annotated	p.value
GO	microtubule plus-end binding (GO:0051010)	GO_MF	2	9	6.00E-05
Disease/ Phenotype	abnormal depression-related behaviour	MPO	2	15	1.50E-04
GO	negative regulation of microtubule depolymerization (GO:0007026)	GO_BP	2	15	1.50E-04
Disease/ Phenotype	small stomach	MPO	2	17	1.80E-04
GO	establishment or maintenance of cell polarity (GO:0007163)	GO_BP	2	18	2.10E-04
Disease/ Phenotype	blepharoptosis	MPO	2	18	2.10E-04

TABLE A.33 ENRICHMENT RESULTS FOR COMBINED RANK TOP 20 GENES IN STAGE 4 GRN INFERRED FROM WANG DATASET USING THE NOVEL Z-SCORE INFERENCE METHOD

Class	Set Name	Source_DB	Annotated genes	Non-annotated	p.value
Pathway	Insulin Signalling	SMPDB	3	35	7.40E-06
Pathway	TNFalpha Signaling Pathway	NetPath	5	334	1.80E-05
Disease/ Phenotype	decreased percent body fat	MPO	3	54	2.50E-05
Pathway	mTOR signaling pathway	PID	3	64	4.10E-05
Pathway	IL1 Signaling Pathway	NetPath	3	65	4.30E-05
Pathway	IL 4 signaling pathway	Biocarta	2	9	5.40E-05
GO	insulin-like growth factor receptor signaling pathway (GO:0048009)	GO_BP	2	10	6.40E-05
Disease/ Phenotype	decreased white adipose tissue amount	MPO	3	77	7.00E-05
GO	positive regulation of glycogen biosynthetic process (GO:0045725)	GO_BP	2	11	7.60E-05
GO	positive regulation of blood vessel endothelial cell migration (GO:0043536)	GO_BP	2	12	8.90E-05

TABLE A.34 ENRICHMENT RESULTS FOR COMBINED RANK TOP 20 GENES IN STAGE 4M GRN INFERRED FROM WANG DATASET USING THE NOVEL Z-SCORE INFERENCE METHOD

Class	Set Name	Source_ DB	Annotated genes	Non-annotated	p. value
Pathway	Purine ribonucleoside monophosphate biosynthesis	Reactome	4	7	3.20E-10
Pathway	purine nucleotides *de novo* biosynthesis II	HumanCyc	4	7	3.20E-10
GO	purine ribonucleoside monophosphate biosynthetic process (GO:0009168)	GO_BP	4	10	9.50E-10
GO	purine nucleotide biosynthetic process (GO:0006164)	GO_BP	4	14	2.90E-09
Pathway	Purine metabolism	Reactome	4	28	3.40E-08
GO	purine base metabolic process (GO:0006144)	GO_BP	4	30	4.40E-08
GO	nucleobase-containing small molecule	GO_BP	4	66	8.40E-07

	metabolic process (GO:0055086)				
Pathway	Metabolism of nucleotides	Reactome	4	66	8.40E-07
Drug/ Chemical	glutamin(CID000000738)	STITCH	3	23	2.70E-06
Pathway	Purine_nucleotides_nucleosides_ metabolism	INOH	4	101	4.30E-06

TABLE A.35 RANKED RETINOBLASTOMA GENES GENERATED BY GÉNIE

Gene	Génie Rank	Score	Gene	Génie Rank	Score	Gene	Génie Rank	Score
RB1	1	500	TP73	41	460	IGF2	81	420
CDKN2A	2	499	CDH1	42	459	EWSR1	82	419
TP53	3	498	SMAD4	43	458	HIF1A	83	418
PTEN	4	497	E2F4	44	457	TFDP1	84	417
E2F1	5	496	MGMT	45	456	MIR34A	85	416
MYCN	6	495	RET	46	455	GSTM1	86	415
EGFR	7	494	HDAC1	47	454	BCL6	87	414
CCND1	8	493	NF2	48	453	ETV6	88	413
BRCA1	9	492	PTCH1	49	452	CDKN2C	89	412
MSH2	10	491	BIRC5	50	451	RAD51	90	411
RBL2	11	490	CDK2	51	450	GSTP1	91	410
MLH1	12	489	CDK6	52	449	MLL	92	409
KRAS	13	488	AKT1	53	448	GSTT1	93	408
ATM	14	487	PMS2	54	447	NDP	94	407
CDK4	15	486	SDHD	55	446	CCND2	95	406
BRCA2	16	485	E2F3	56	445	TSC2	96	405
RBL1	17	484	FGFR3	57	444	IDH1	97	404
RASSF1	18	483	CCNE1	58	443	SRY	98	403
FHIT	19	482	STK11	59	442	PHOX2B	99	402
MDM2	20	481	MSH6	60	441	MECP2	100	401
VHL	21	480	AURKA	61	440	TET2	101	400
BRAF	22	479	HRAS	62	439	ID1	102	399
CDKN1B	23	478	TP63	63	438	CCNA2	103	398
APC	24	477	HMGA2	64	437	RUNX3	104	397
KIT	25	476	FLCN	65	436	MKI67	105	396
WT1	26	475	GJB2	66	435	CHEK2	106	395
SMARCB1	27	474	WWOX	67	434	ING1	107	394
MEN1	28	473	PROM1	68	433	CD44	108	393
CDKN1A	29	472	PAX6	69	432	ID2	109	392
ERBB2	30	471	BCL2	70	431	DMBT1	110	391
MYC	31	470	FGFR2	71	430	MYBL2	111	390
NF1	32	469	RUNX1	72	429	DLC1	112	389
TERT	33	468	SP1	73	428	KLF6	113	388
CTNNB1	34	467	CCND3	74	427	NOTCH1	114	387
ALK	35	466	AR	75	426	RBBP4	115	386
PIK3CA	36	465	E2F5	76	425	TFE3	116	385
PDGFRA	37	464	TOP2A	77	424	BAX	117	384
NBN	38	463	E2F2	78	423	BMI1	118	383
PRDM2	39	462	FGFR4	79	422	XRCC1	119	382
CDKN2B	40	461	MTAP	80	421	VEGFA	120	381

Gene	Génie Rank	Score	Gene	Génie Rank	Score	Gene	Génie Rank	Score
EXT1	121	380	PRAME	170	331	DNMT1	219	282
SMARCA4	122	379	CADM1	171	330	PLK1	220	281
ERG	123	378	TGFBI	172	329	PTGS2	221	280
H2AFX	124	377	DCC	173	328	FGFR1	222	279
JAK2	125	376	DAZ1	174	327	AURKB	223	278
CHM	126	375	ZEB2	175	326	EP300	224	277
CYLD	127	374	LGALS3	176	325	PTPRD	225	276
EZH2	128	373	RPGR	177	324	EZR	226	275
TYMS	129	372	FBXW7	178	323	FAM123B	227	274
NME1	130	371	TPD52	179	322	GJB6	228	273
SOX2	131	370	HIC1	180	321	TMPRSS2	229	272
EXT2	132	369	CD24	181	320	ABCB1	230	271
ABCC6	133	368	PTTG1	182	319	DAZ4	231	270
MUTYH	134	367	HPRT1	183	318	UBE3A	232	269
TFAP2A	135	366	PLAGL1	184	317	CYP1B1	233	268
PRPF31	136	365	PAX8	185	316	DBC1	234	267
FAS	137	364	NKX3-1	186	315	XRCC3	235	266
SHOX	138	363	DMD	187	314	DLK1	236	265
PAX5	139	362	FOXO1	188	313	LZTS1	237	264
PARK2	140	361	AZF1	189	312	POU5F1	238	263
PAX2	141	360	NQO1	190	311	CYP21A2	239	262
LMNA	142	359	HDAC2	191	310	MAPK1	240	261
SMN1	143	358	PRKAR1A	192	309	FH	241	260
CAV1	144	357	NRAS	193	308	RPE65	242	259
SDHB	145	356	CASP8	194	307	PML	243	258
FOXC1	146	355	SRPX	195	306	NGFR	244	257
ARID4A	147	354	NPM1	196	305	BEST1	245	256
MTOR	148	353	STAT3	197	304	SKP2	246	255
PCNA	149	352	USP9Y	198	303	NTRK3	247	254
TNF	150	351	CDKN1C	199	302	CASZ1	248	253
ABCA4	151	350	GPC3	200	301	NDRG1	249	252
OPA1	152	349	TIMP3	201	300	JAK3	250	251
ST7	153	348	ATR	202	299	CACNA1F	251	250
TGFB1	154	347	ERCC2	203	298	SPAST	252	249
NPHP1	155	346	MDM4	204	297	DKK3	253	248
PHB	156	345	IL24	205	296	CDKN2D	254	247
NFKB1	157	344	PMP22	206	295	MAGEC2	255	246
TSPY1	158	343	IKZF1	207	294	TGFBR2	256	245
NSD1	159	342	XPC	208	293	SMARCA2	257	244
RBBP7	160	341	SHH	209	292	TFDP2	258	243
PAX3	161	340	SFN	210	291	GNAS	259	242
NTRK1	162	339	CTAG1B	211	290	FGF2	260	241
MET	163	338	PDPN	212	289	TNFSF10	261	240
TSC1	164	337	NTRK2	213	288	MRE11A	262	239
NKX2-1	165	336	CEBPA	214	287	CTDSPL	263	238
ESR1	166	335	AXIN2	215	286	PRDM1	264	237
AIP	167	334	ABL1	216	285	OCA2	265	236
ERCC1	168	333	XRCC5	217	284	PRKDC	266	235
RS1	169	332	CA9	218	283	MAGEA1	267	234

Gene	**Génie Rank**	**Score**	**Gene**	**Génie Rank**	**Score**	**Gene**	**Génie Rank**	**Score**
RECQL4	268	233	BLM	317	184	SSX4	366	135
ABCD1	269	232	DEK	318	183	SNCG	367	134
TWIST1	270	231	CTCF	319	182	CBL	368	133
DAZ3	271	230	NES	320	181	TMPO	369	132
MAD2L1	272	229	NRG1	321	180	APAF1	370	131
IGF1R	273	228	ATP7B	322	179	RAF1	371	130
EPAS1	274	227	SLC26A4	323	178	LTA	372	129
DAPK1	275	226	RCVRN	324	177	ESR2	373	128
SERPINB5	276	225	EYA1	325	176	MMP1	374	127
CYP1A1	277	224	KLF4	326	175	KRT19	375	126
MAGEA3	278	223	SSX2	327	174	IGFBP5	376	125
DAZ2	279	222	EID1	328	173	RPS19	377	124
FOXL2	280	221	BNIP3	329	172	FOXP1	378	123
MIR21	281	220	TSG101	330	171	ZFHX3	379	122
HNF1B	282	219	LPAR6	331	170	MAGEA4	380	121
AHR	283	218	OTX2	332	169	MIR15A	381	120
TCF4	284	217	PITX2	333	168	CEBPB	382	119
MCL1	285	216	TNFRSF10A	334	167	TNFRSF10B	383	118
GJA1	286	215	CXCR4	335	166	BUB1	384	117
USH2A	287	214	CASP3	336	165	PARK7	385	116
MITF	288	213	EPB41L3	337	164	MXI1	386	115
JAG1	289	212	DLEU2	338	163	NDN	387	114
POLH	290	211	TNFRSF1A	339	162	PKHD1	388	113
OGG1	291	210	DUSP6	340	161	PDGFB	389	112
TUSC2	292	209	MUC1	341	160	GLI1	390	111
TES	293	208	SERPINF1	342	159	ID4	391	110
KDM5A	294	207	GATA3	343	158	FZD4	392	109
XIAP	295	206	XPA	344	157	FANCA	393	108
IKBKG	296	205	FBN1	345	156	TGFBR1	394	107
RBM5	297	204	SOX11	346	155	COL2A1	395	106
ABCC1	298	203	MDK	347	154	KIF1B	396	105
S100A4	299	202	PKD1	348	153	LIMD1	397	104
GPR143	300	201	DCLRE1C	349	152	FOXO3	398	103
NPRL2	301	200	COL4A5	350	151	PTPN11	399	102
DACH1	302	199	EBF3	351	150	THBS1	400	101
EBAG9	303	198	NANOG	352	149	IGH@	401	100
MAGEC1	304	197	ELOVL4	353	148	EGF	402	99
TRIM13	305	196	EPHB2	354	147	ZMYND10	403	98
PLAG1	306	195	PROX1	355	146	SOX9	404	97
PSMD10	307	194	MTHFR	356	145	ABCC8	405	96
RRM2B	308	193	MYOC	357	144	HMGA1	406	95
NCAM1	309	192	RBBP9	358	143	MIR16-1	407	94
CHD5	310	191	MLL5	359	142	ASPSCR1	408	93
CDC73	311	190	FLT3	360	141	SP3	409	92
CLN3	312	189	RBBP8	361	140	NFE2L2	410	91
BSG	313	188	CHFR	362	139	MYB	411	90
KISS1	314	187	CREBBP	363	138	NOTCH3	412	89
KLK10	315	186	FANCD2	364	137	HDAC3	413	88
GDAP1	316	185	GAGE1	365	136	BIN1	414	87

Gene	Génie Rank	Score	Gene	Génie Rank	Score	Gene	Génie Rank	Score
CKS1B	415	86	WNT1	444	57	HLA-G	473	28
PPP2R1B	416	85	SLC2A1	445	56	SMAD5	474	27
RUNX2	417	84	SOD2	446	55	TYR	475	26
SSX1	418	83	HBP1	447	54	DICER1	476	25
CAMTA1	419	82	ASCL1	448	53	PYCARD	477	24
EGR1	420	81	MTDH	449	52	HGF	478	23
TERC	421	80	MTA1	450	51	NEUROD1	479	22
YAP1	422	79	ZIC2	451	50	TP53BP1	480	21
ILK	423	78	FRMD7	452	49	MMP2	481	20
CCNDBP1	424	77	MERTK	453	48	EPHA2	482	19
PRPH2	425	76	HSPA5	454	47	FGF14	483	18
FOXM1	426	75	CYR61	455	46	PPARG	484	17
GADD45A	427	74	COL18A1	456	45	MSH3	485	16
SFRP1	428	73	ARNT	457	44	UMOD	486	15
CDK1	429	72	CDC25A	458	43	USH1C	487	14
PURA	430	71	MYOD1	459	42	REG4	488	13
L3MBTL1	431	70	JUN	460	41	CLDN4	489	12
PTK2	432	69	MAPK3	461	40	PIK3CG	490	11
BCR	433	68	SIX3	462	39	SCN1A	491	10
SPARC	434	67	SSX4B	463	38	AAAS	492	9
NOD2	435	66	RASSF2	464	37	BACH2	493	8
CD82	436	65	KCNQ1OT1	465	36	TAL1	494	7
LMO1	437	64	FIP1L1	466	35	BRIP1	495	6
UVRAG	438	63	AMACR	467	34	XAF1	496	5
LIN9	439	62	BARD1	468	33	CTTN	497	4
NR0B1	440	61	WRN	469	32	RP2	498	3
IGF2BP3	441	60	ESD	470	31	MAX	499	2
C9orf72	442	59	OTC	471	30	PARP1	500	1
SET	443	58	DCN	472	29			

TABLE A.36 ENRICHMENT RESULTS FOR GENES IN THE LARGEST UNIQUE CLIQUE IN THE RB BLUE GRN INFERRED USING POSITIVE Z SCORE METHOD

Class	Set Name	Source_DB	Annotated genes	Non-annotated	p.value
GO	axon guidance (GO:0007411)	GO_BP	11	294	5.70E-06
Disease/ Phenotype	decreased cell proliferation	MPO	10	272	1.80E-05
Disease/ Phenotype	abnormal nervous system development	MPO	5	53	4.00E-05
Pathway	Transport of vitamins, nucleosides, and related molecules	Reactome	4	27	4.90E-05
Pathway	Axon guidance	Reactome	9	257	6.80E-05
Disease/ Phenotype	abnormal spinal nerve morphology	MPO	4	30	7.20E-05
Disease/ Phenotype	complete perinatal lethality	MPO	9	265	8.50E-05

TABLE A.37 ENRICHMENT RESULTS FOR GENES IN THE LARGEST UNIQUE CLIQUE IN THE RB GREEN GRN INFERRED USING POSITIVE Z SCORE METHOD

Class	Set Name	Source_DB	Annotated genes	Non-annotated	p.value
GO	visual perception (GO:0007601)	GO_BP	21	157	1.20E-15
Pathway	Transmission across Chemical Synapses	Reactome	18	172	6.10E-12
Disease/ Phenotype	abnormal eye electrophysiology	MPO	14	104	7.20E-11
GO	synaptic transmission (GO:0007268)	GO_BP	23	366	1.10E-10
Disease/ Phenotype	retinal degeneration	MPO	12	72	1.90E-10
Pathway	Neurotransmitter Receptor Binding And Downstream Transmission In The Postsynaptic Cell	Reactome	14	122	4.90E-10
Pathway	Neuronal System	Reactome	19	264	6.20E-10
Pathway	Visual signal transduction: Rods	PID	7	16	4.70E-09
Drug/Chemical	Kainite (CID000010255)	STITCH	11	97	4.10E-08
GO	nervous system development (GO:0007399)	GO_BP	20	389	4.80E-08

TABLE A.38 ENRICHMENT RESULTS FOR COMBINED RANK TOP 20 GENES IN RB RED GRN INFERRED USING THE POSITIVE Z SCORE METHOD

Class	Set Name	Source_DB	Annotated genes	Non-annotated	p.value
Drug/ Chemical	acetyl coenzyme-A(CID000000181)	STITCH	3	53	2.00E-05
Drug/ Chemical	potassium hydride(CID000082127)	STITCH	3	64	3.50E-05
Pathway	Glycolysis and Gluconeogenesis	EHMN	3	77	5.90E-05
Pathway	2-Hydroxyglutric Aciduria (D And L Form)	SMPDB	2	13	9.20E-05
Drug/ Chemical	acetyl-CoA(CID000006302)	STITCH	3	90	9.30E-05
GO	myosin binding (GO:0017022)	GO_MF	2	14	1.00E-04
Pathway	Pyruvate Metabolism	SMPDB	2	18	1.70E-04
Pathway	Leigh Syndrome	SMPDB	2	18	1.70E-04
Drug/ Chemical	acetyl-CoA(CID100000181)	STITCH	3	121	2.20E-04
GO	positive regulation of protein catabolic process (GO:0045732)	GO_BP	2	21	2.20E-04

TABLE A.39 ENRICHMENT RESULTS FOR COMBINED RANK TOP 20 GENES IN RB GREEN GRN INFERRED USING THE POSITIVE Z SCORE METHOD

Class	Set Name	Source_DB	Annotated genes	Non-annotated	p.value
Pathway	Visual signal transduction: Rods	PID	3	20	1.60E-06
Drug/Chemical	alprazolam(CID000002118)	STITCH	3	34	6.80E-06
Drug/ Chemical	alprazolam(CID100002118)	STITCH	3	35	7.40E-06
Disease/ Phenotype	Memory impairment	SIDER	4	155	1.80E-05
Disease/ Phenotype	Difficulty in micturition	SIDER	3	50	2.00E-05
Disease/ Phenotype	Movements involuntary	SIDER	3	50	2.00E-05
Disease/ Phenotype	Abdominal distress	SIDER	3	50	2.00E-05
Disease/ Phenotype	Dysarthria	SIDER	3	50	2.00E-05
Disease/ Phenotype	Abnormal involuntary movements	SIDER	3	50	2.00E-05
Disease/ Phenotype	Cognitive disorder	SIDER	3	50	2.00E-05

TABLE A.40 ENRICHMENT RESULTS FOR THE GENES IN LARGEST UNIQUE CLIQUE IN THE RB BLUE GRN INFERRED USING NEGATIVE Z SCORE METHOD

Class	Set Name	Source_DB	Annotated genes	Non-annotated	p.value
GO	helicase activity (GO:0004386)	GO_MF	7	112	2.20E-06
Drug/Chemical	fulvestrant	CTD	9	242	4.50E-06
GO	translation (GO:0006412)	GO_BP	9	258	7.40E-06
GO	nuclear mRNA splicing, via spliceosome (GO:0000398)	GO_BP	7	160	2.00E-05
Disease/ Phenotype	complete embryonic lethality before implantation	MPO	6	118	3.60E-05
Pathway	Processing of Capped Intron-Containing Pre-mRNA	Reactome	6	131	6.30E-05
Pathway	Translation Factors(WP107)	WikiPathways	4	46	1.10E-04
Pathway	mRNA Processing	Reactome	6	150	1.30E-04
Disease/ Phenotype	Classification of acute myeloid leukemias	MethyCancer	5	93	1.30E-04
GO	translation initiation factor activity (GO:0003743)	GO_MF	4	50	1.50E-04

TABLE A.41 ENRICHMENT RESULTS FOR THE GENES IN LARGEST UNIQUE CLIQUE IN THE RB RED GRN INFERRED USING NEGATIVE Z SCORE METHOD

Class	Set Name	Source_DB	Annotated genes	Non-annotated	p.value
GO	keratinization (GO:0031424)	GO_BP	10	33	2.50E-14
GO	keratin filament (GO:0045095)	GO_CC	10	83	8.40E-11
GO	cytokine activity (GO:0005125)	GO_MF	7	156	3.10E-05

TABLE A.42 ENRICHMENT RESULTS FOR THE GENES IN LARGEST UNIQUE CLIQUE IN THE RB GREEN GRN INFERRED USING NEGATIVE Z SCORE METHOD

Class	Set Name	Source_DB	Annotated genes	Non-annotated	p.value
Drug/Chemical	Lucanthone	CTD	54	159	1.60E-69
Drug/Chemical	dasatinib	CTD	59	406	2.90E-57
Drug/Chemical	Polychlorinated Biphenyls	CTD	38	155	1.20E-43
Pathway	Cell Cycle	Reactome	42	367	5.20E-36
GO	mitotic cell cycle (GO:0000278)	GO_BP	38	280	5.20E-35
Pathway	Cell Cycle, Mitotic	Reactome	37	293	5.00E-33
Drug/Chemical	trans-10,cis-12-conjugated linoleic acid	CTD	28	116	6.10E-32
Pathway	DNA Replication	Reactome	30	170	1.30E-30
GO	cell cycle (GO:0007049)	GO_BP	35	399	2.90E-26
GO	DNA replication (GO:0006260)	GO_BP	23	127	8.20E-24

TABLE A.43 ENRICHMENT RESULTS FOR COMBINED RANK TOP 20 GENES IN RB GREEN GRN INFERRED USING THE NEGATIVE Z SCORE METHOD

Class	Set Name	Source_DB	Annotated genes	Non-annotated	p.value
GO	mitotic cell cycle (GO:0000278)	GO_BP	9	309	2.30E-12
Pathway	DNA Replication	Reactome	8	192	3.40E-12
Pathway	Cell Cycle	Reactome	9	400	2.20E-11
Pathway	Mitotic M-M/G1 phases	Reactome	7	171	1.10E-10
Pathway	Cell Cycle, Mitotic	Reactome	8	322	1.90E-10
GO	mitotic prometaphase (GO:0000236)	GO_BP	4	80	8.70E-07
GO	M phase of mitotic cell cycle (GO:0000087)	GO_BP	4	88	1.30E-06
Pathway	Mitotic Prometaphase	Reactome	4	88	1.30E-06
Pathway	M Phase	Reactome	4	92	1.50E-06
Drug/Chemical	dasatinib	CTD	6	459	2.30E-06
GO	DNA strand elongation involved in DNA replication (GO:0006271)	GO_BP	3	28	2.80E-06
Pathway	DNA strand elongation	Reactome	3	28	2.80E-06
Disease/ Phenotype	abnormal cell cycle	MPO	4	116	3.60E-06
Drug/Chemical	trans-10,cis-12-conjugated linoleic acid	CTD	4	140	7.50E-06
GO	DNA replication (GO:0006260)	GO_BP	4	146	8.80E-06
GeneRegulation	E2F1	TFactS	4	158	1.20E-05
Disease/ Phenotype	increased tumor incidence	MPO	4	159	1.20E-05
Drug/ Chemical	Polychlorinated Biphenyls	CTD	4	189	2.40E-05
GO	condensed chromosome kinetochore (GO:0000777)	GO_CC	3	60	2.40E-05
Disease/ Phenotype	sarcoma	MPO	3	64	2.90E-05

APPENDIX B

VENN DIAGRAMS OF GENE DISTRIBUTION FOR THE VARIOUS GRNS INFERRED FROM THE DIFFERENT MICROARRAY DATASETS

The following appendix contains Venn diagrams of the distribution of genes in both the largest cliques, and also the top 20 ranked genes for the combined metric for the various GRNs in the work.

FIGURE B.1 VENN DIAGRAM OF UNIQUE AND COMMON GENES IN THE CLIQUES IN THE GLIOBLASTOMA CATEGORY GRNS INFERRED USING WGCNA

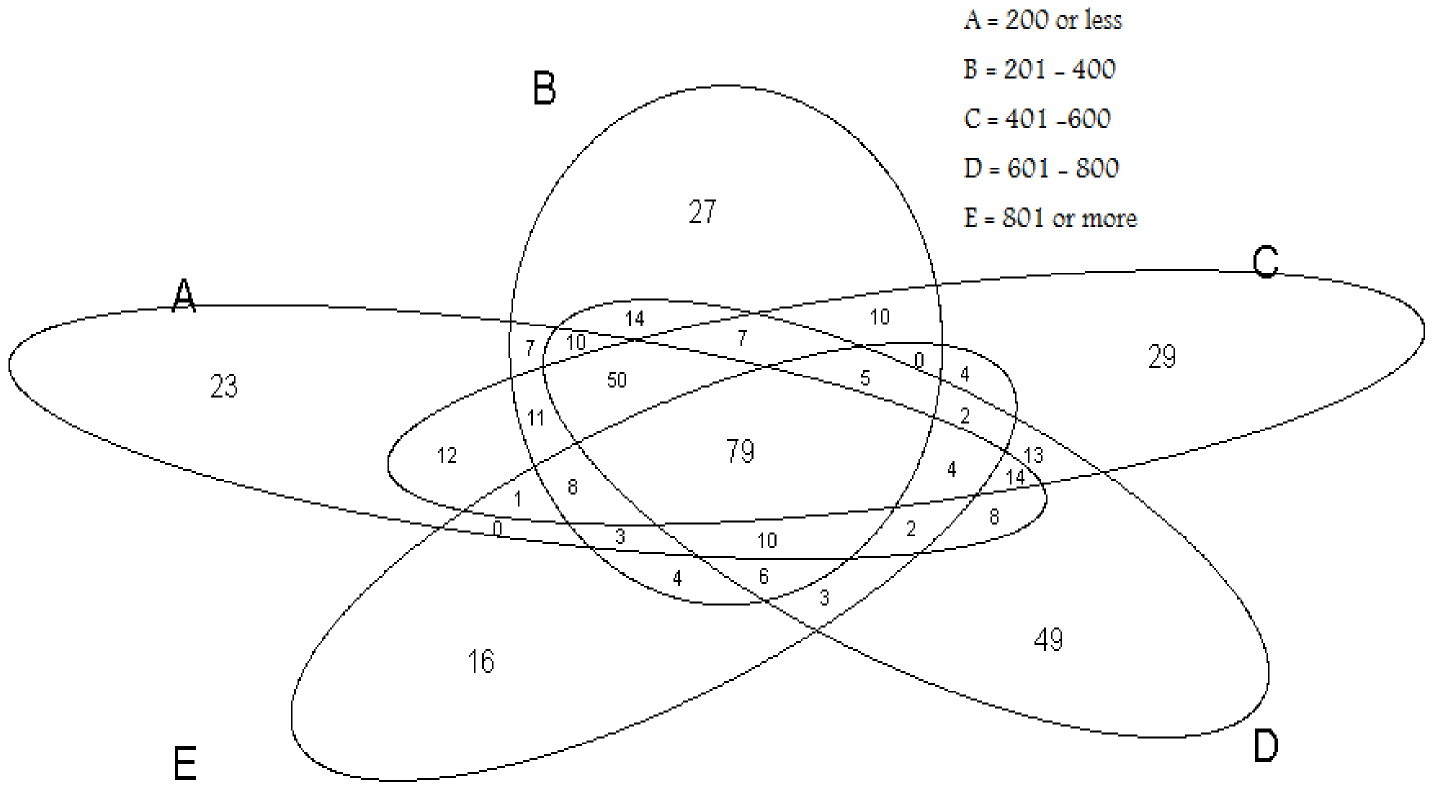


FIGURE B.2 VENN DIAGRAM OF UNIQUE AND COMMON GENES IN THE TOP 20 RANKING GENES FOR EACH METRIC IN THE 0-200 GLIOBLASTOMA GRN INFERRED USING WGCNA

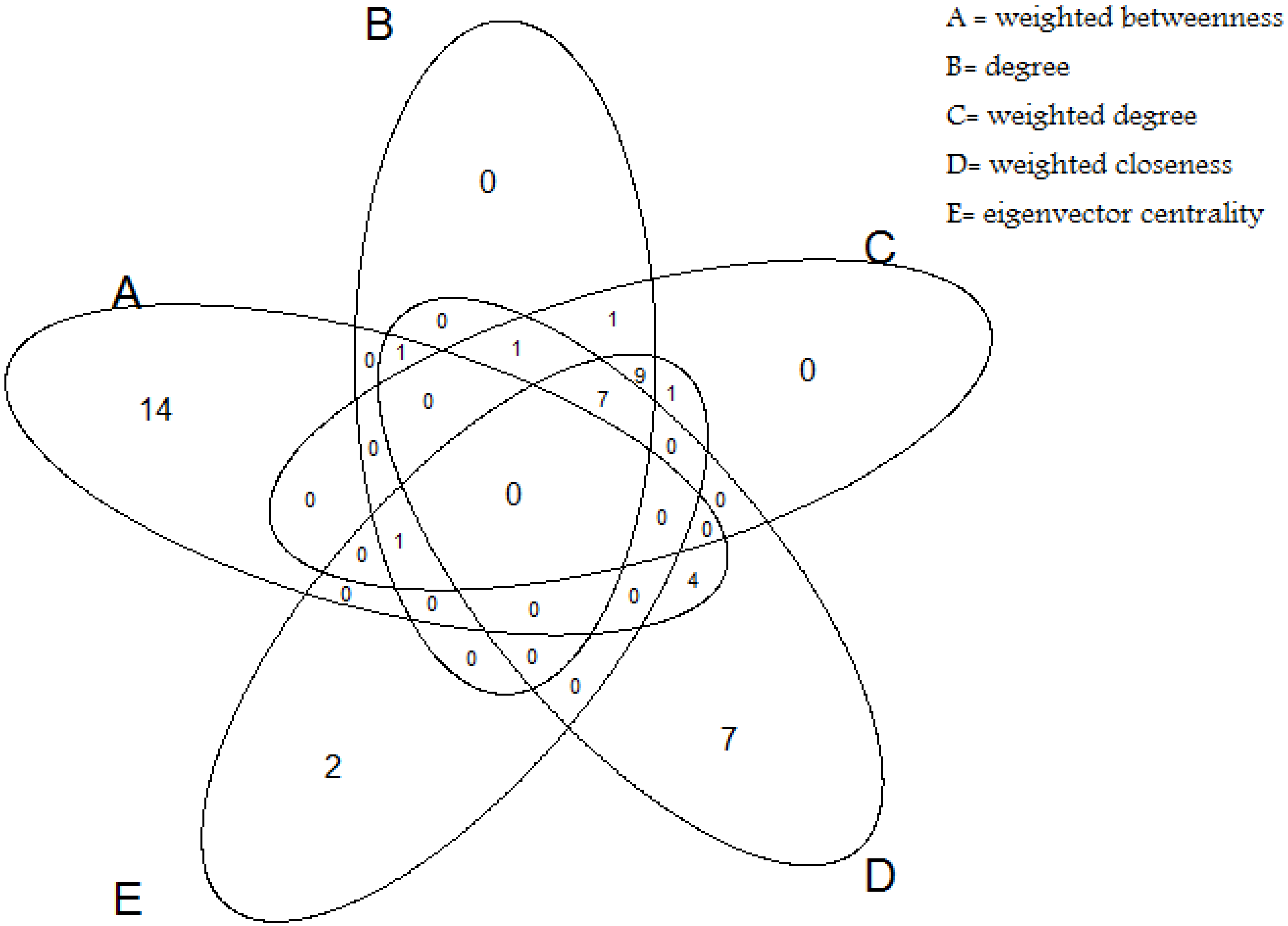


FIGURE B.3 VENN DIAGRAM OF UNIQUE AND COMMON GENES IN THE LARGEST UNIQUE CLIQUES IN THE DIFFERENT CATEGORIES OF GLIOBLASTOMA GRN INFERRED USING THE NOVEL INFERENCE METHOD

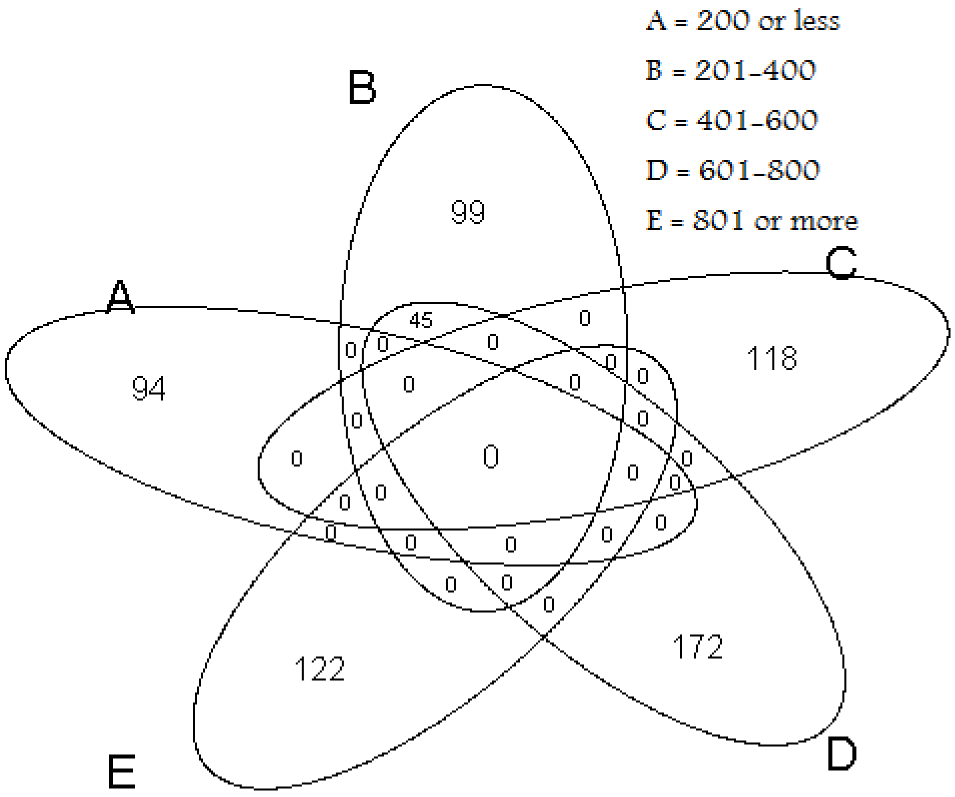


FIGURE B.4 VENN DIAGRAM OF GENE DISTRIBUTION IN THE TOP 20 COMBINED METRIC RANKINGS OF THE GLIOBLASTOMA CATEGORY GRNS INFERRED USING NOVEL INFERENCE METHOD

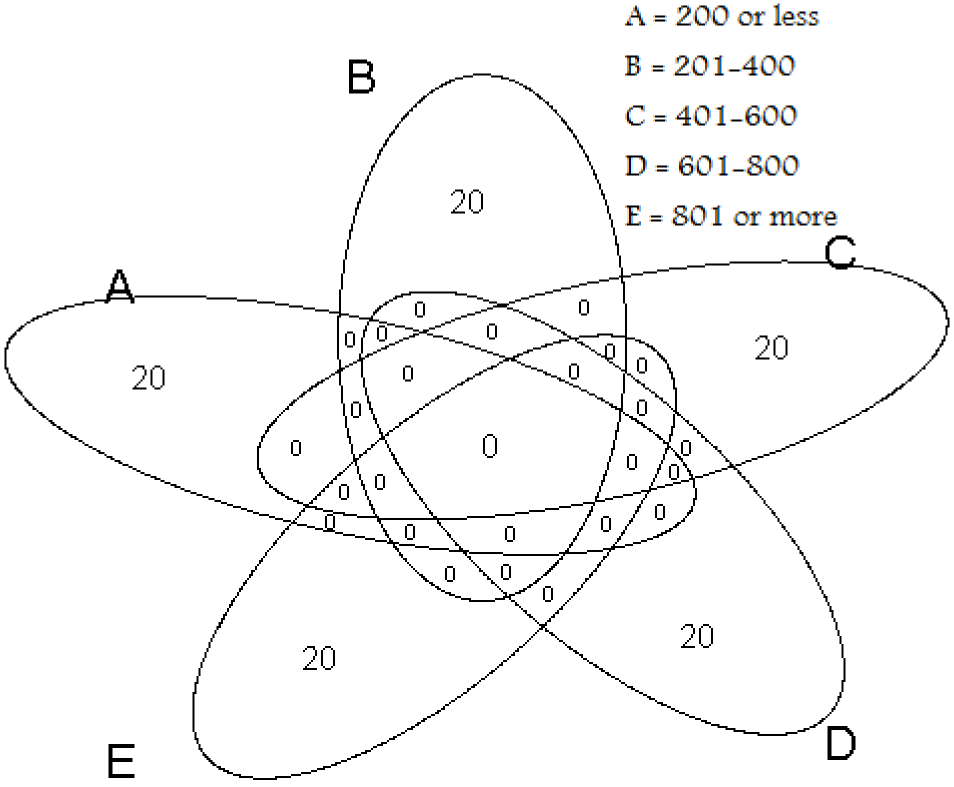


FIGURE B.5 VENN DIAGRAM OF GENE DISTRIBUTION IN THE TOP 100 COMBINED METRIC RANKINGS OF THE GLIOBLASTOMA CATEGORY GRNS INFERRED USING NOVEL INFERENCE METHOD

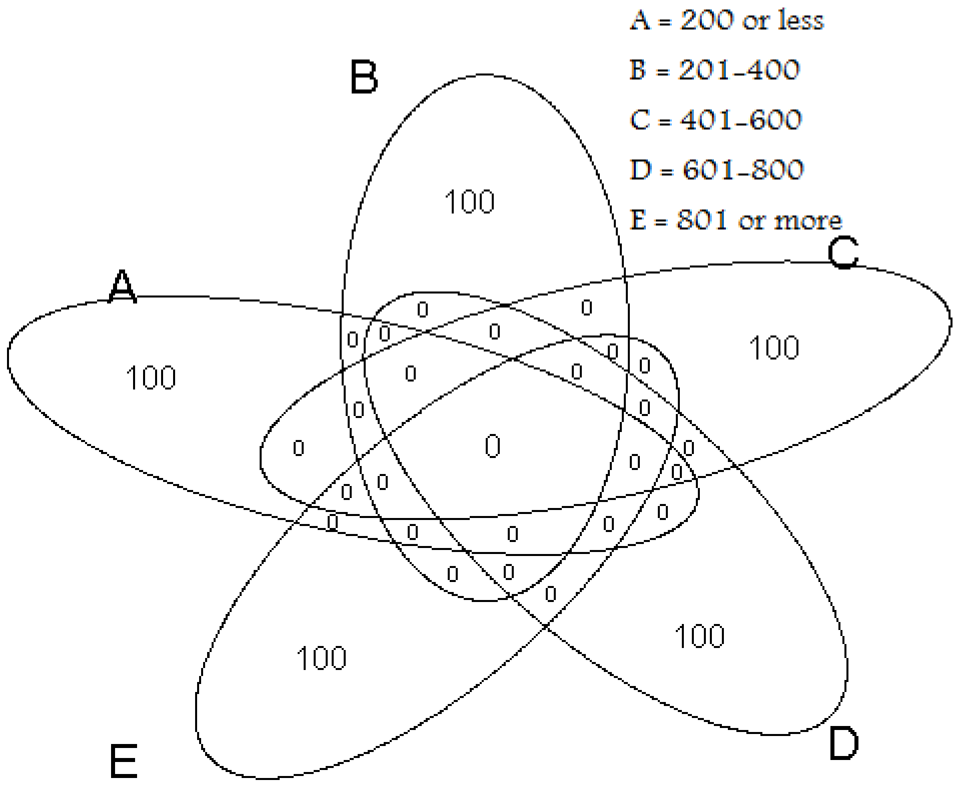


FIGURE B.6 VENN DIAGRAM OF GENE DISTRIBUTION IN THE LARGEST UNIQUE CLIQUES IN THE GRNS INFERRED FROM THE MOLENAAR DATASET USING NOVEL INFERENCE METHOD

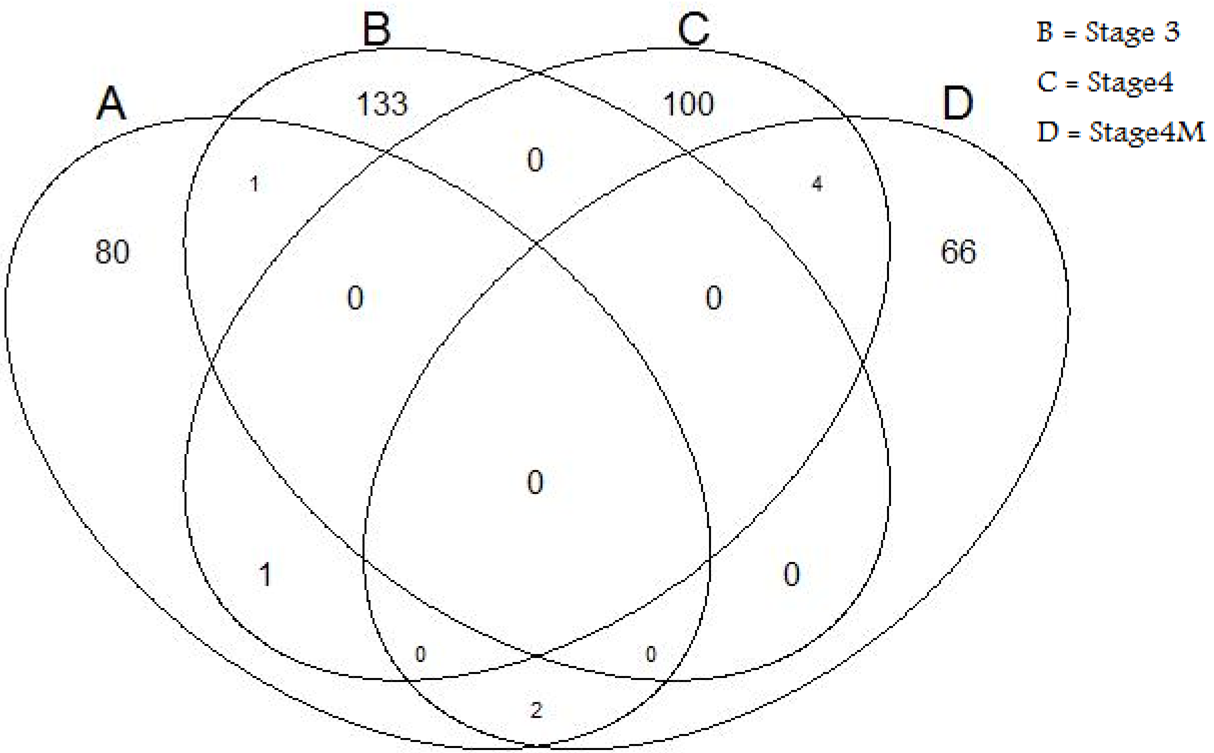


FIGURE B.7 VENN DIAGRAM OF GENE DISTRIBUTION IN THE LARGEST UNIQUE CLIQUES IN THE GRNS INFERRED FROM THE WANG DATASET USING NOVEL INFERENCE METHOD

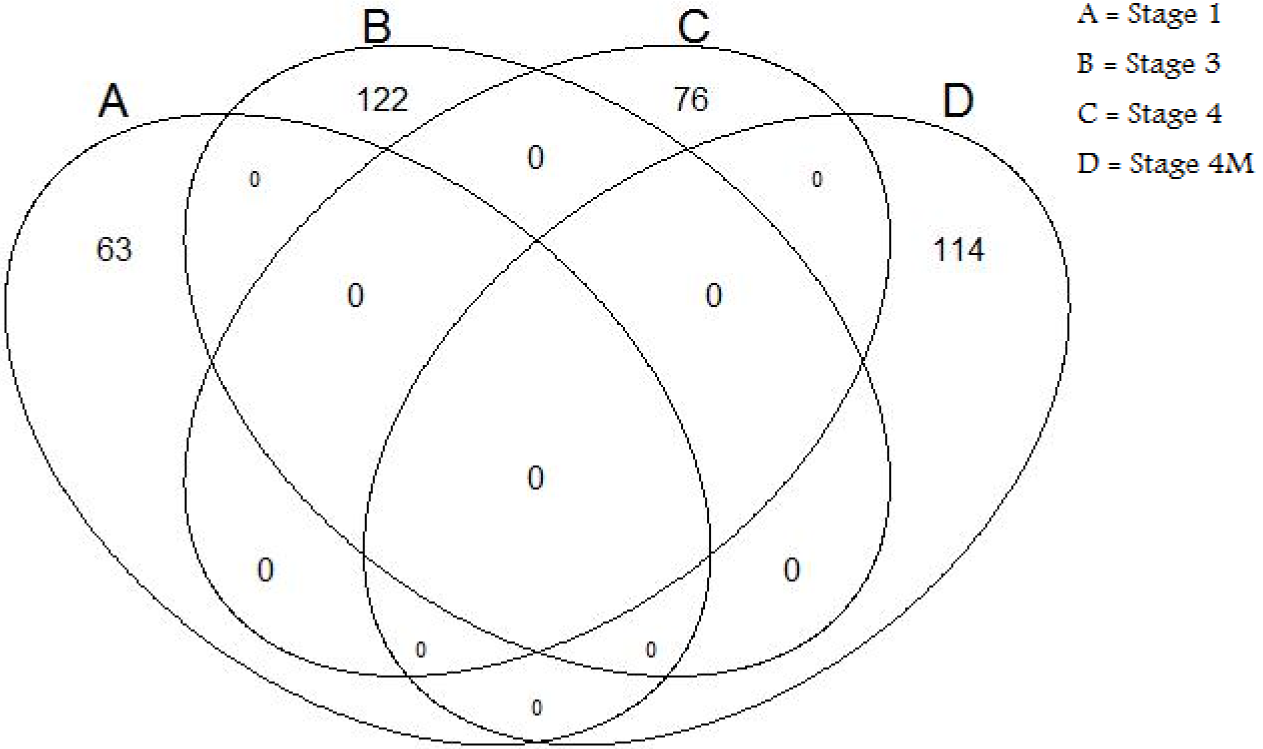


FIGURE B.8 GENE DISTRIBUTION OF TOP 20 COMBINED METRIC RANKINGS IN THE GRNS INFERRED FROM THE MOLENAAR DATASET

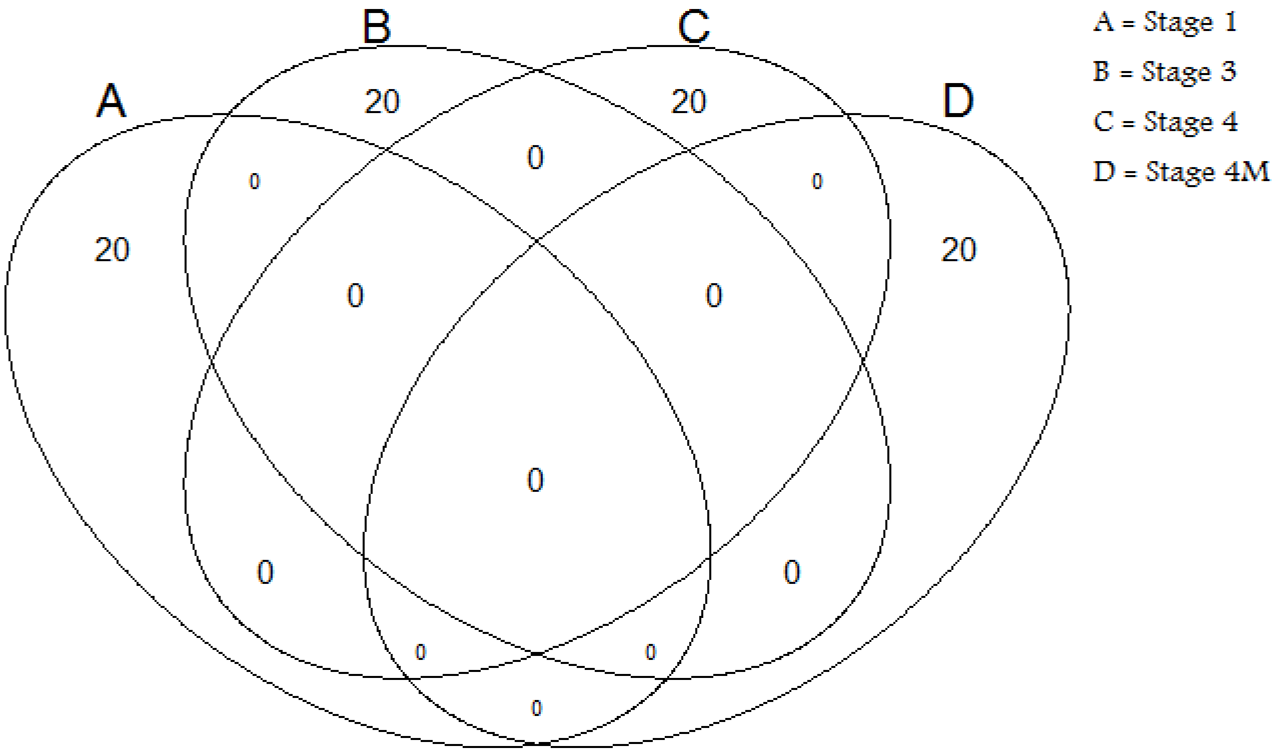
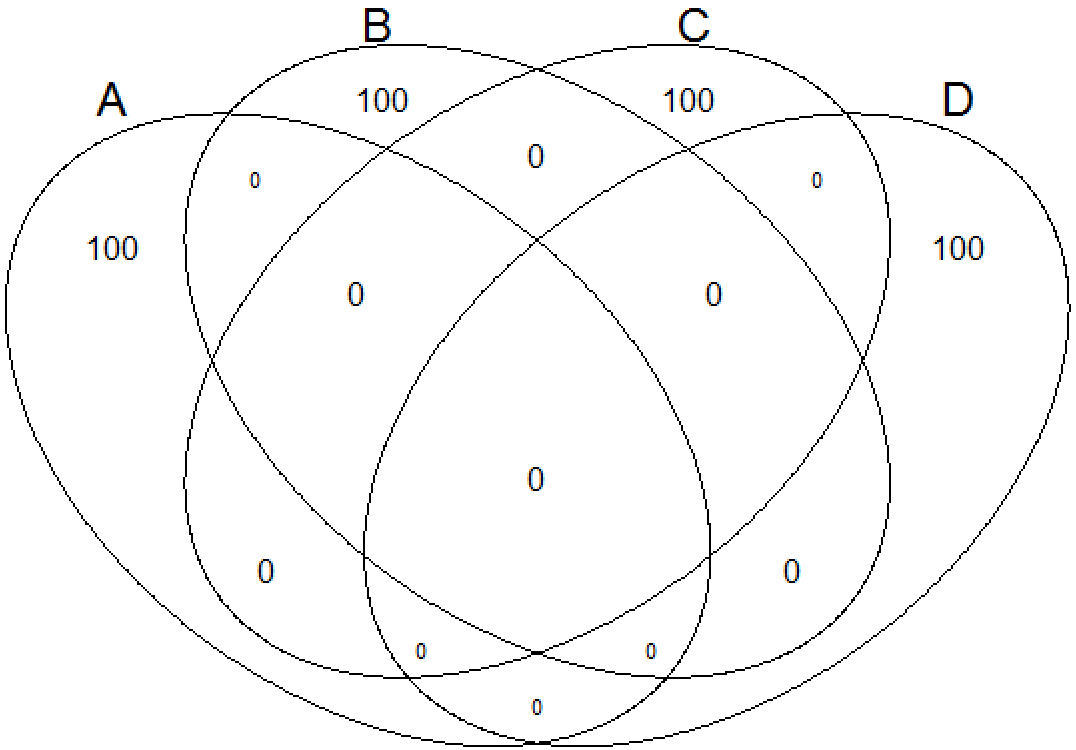


FIGURE B.9 GENE DISTRIBUTION OF TOP 100 COMBINED METRIC RANKINGS IN THE GRNS INFERRED FROM THE MOLENAAR DATASET



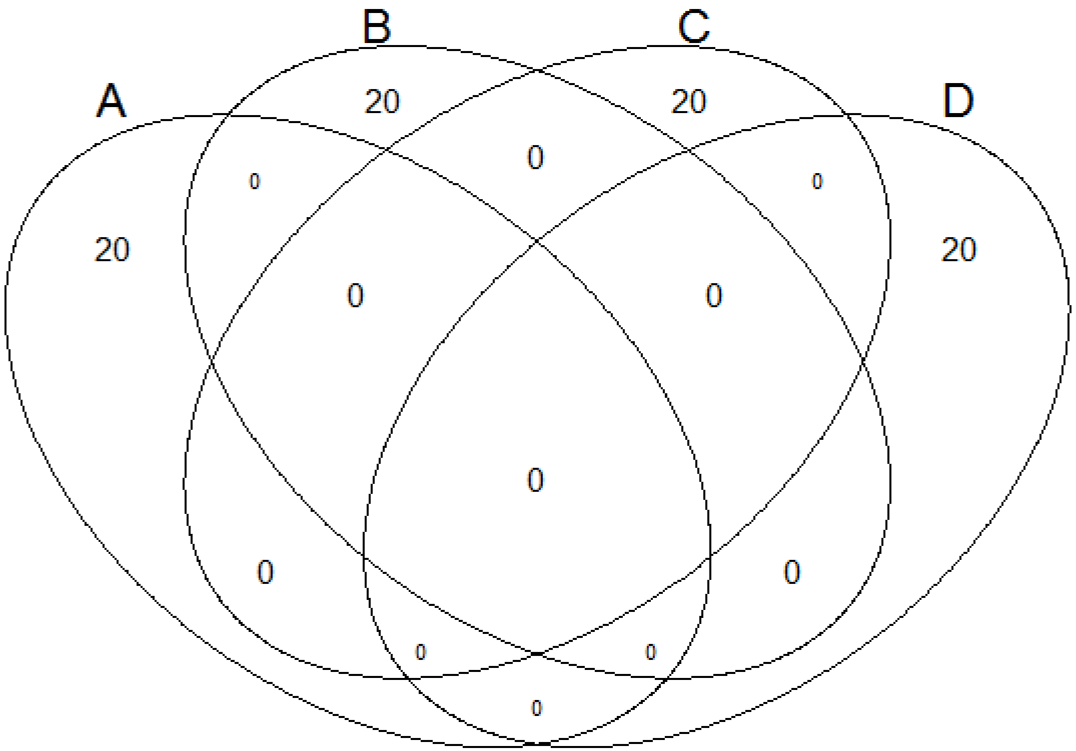
A = Stage 1

B = Stage 3

C = Stage 4

D = Stage 4M

FIGURE B.10 GENE DISTRIBUTION OF TOP 20 COMBINED METRIC RANKINGS IN THE GRNS INFERRED FROM THE WANG DATASET



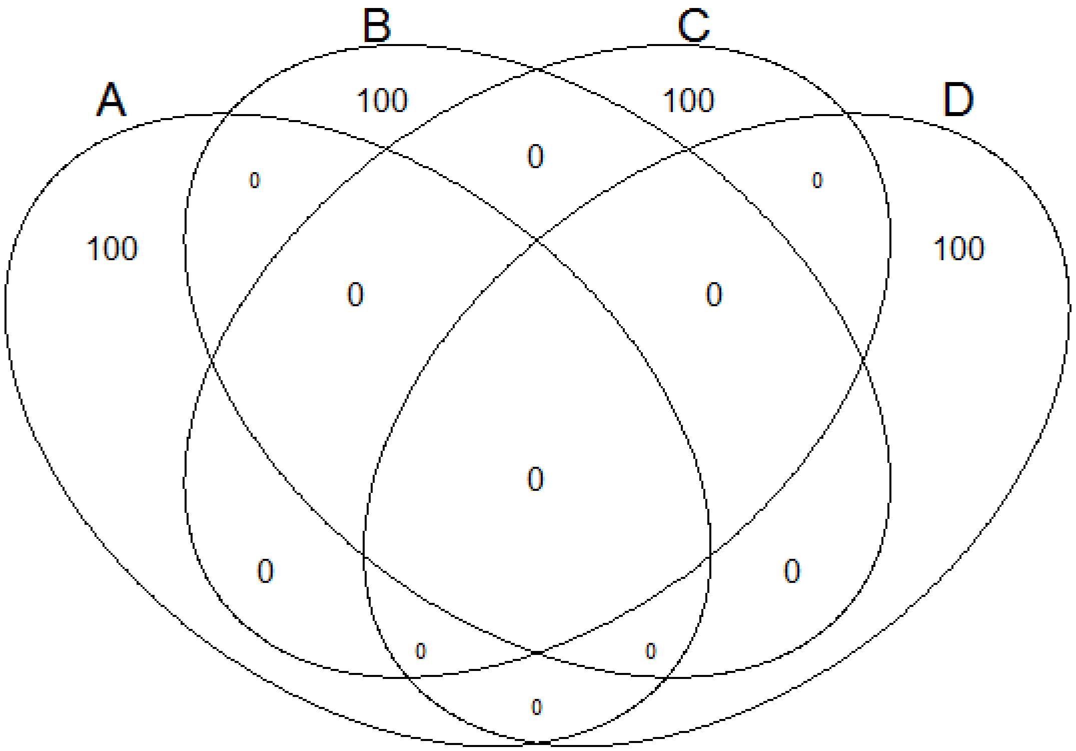
A = Stage 1

B = Stage 3

C = Stage 4

D = Stage 4M

FIGURE B.11 GENE DISTRIBUTION OF TOP 100 COMBINED METRIC RANKINGS IN THE GRNS INFERRED FROM THE WANG DATASET



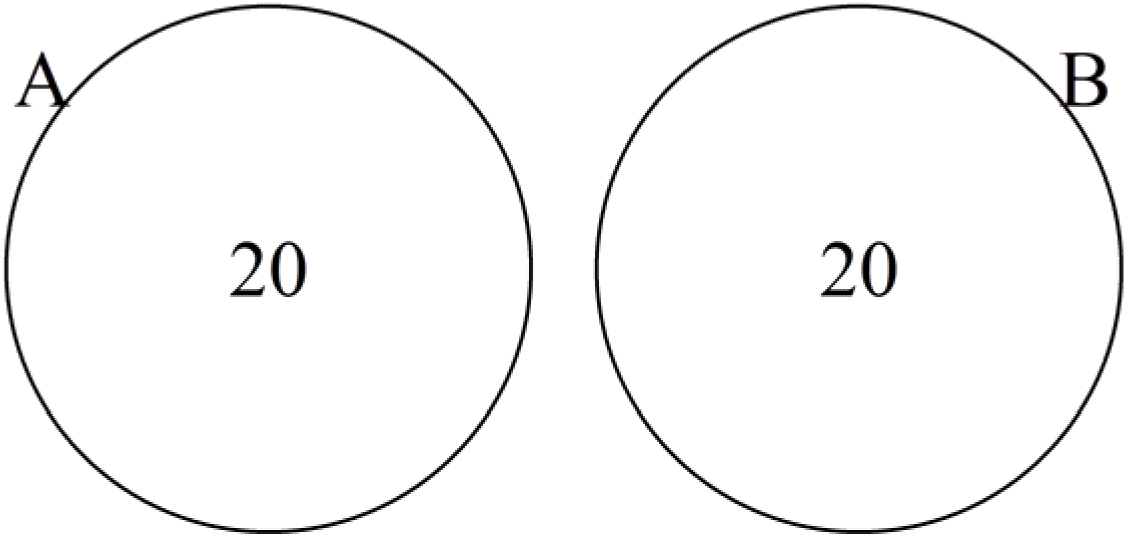
A = Stage 1

B = Stage 3

C = Stage 4

D = Stage 4M

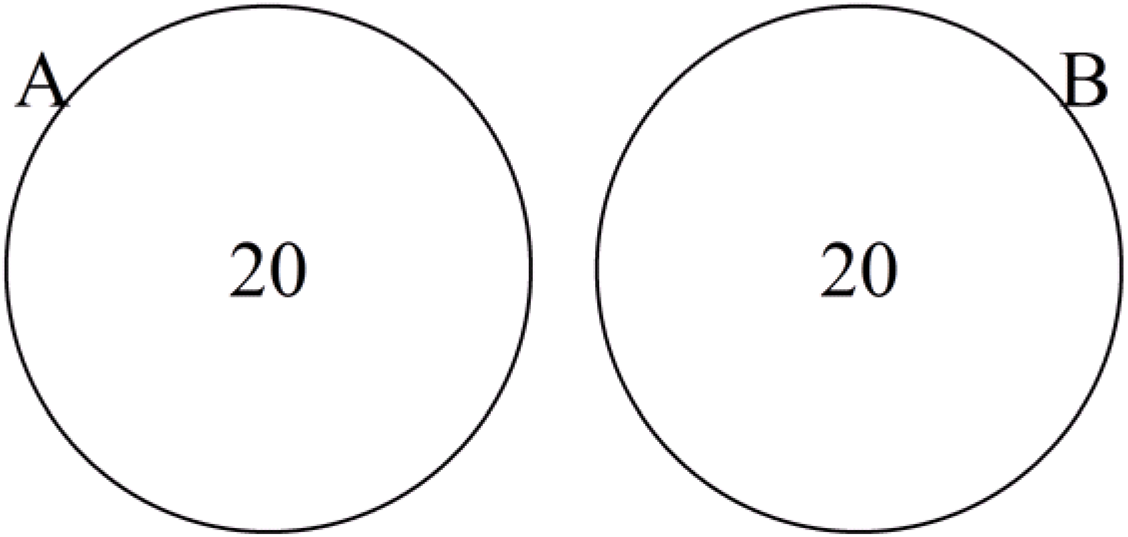
FIGURE B.12 VENN DIAGRAM STAGE1 TOP 20 RANKED GENES MOLENAAR AND WANG DATASETS



A = MOLENAAR STAGE 1 GRN TOP 20 RANKED GENES FOR COMBINED METRIC

B = WANG STAGE 1 GRN TOP 20 RANKED GENES FOR COMBINED METRIC

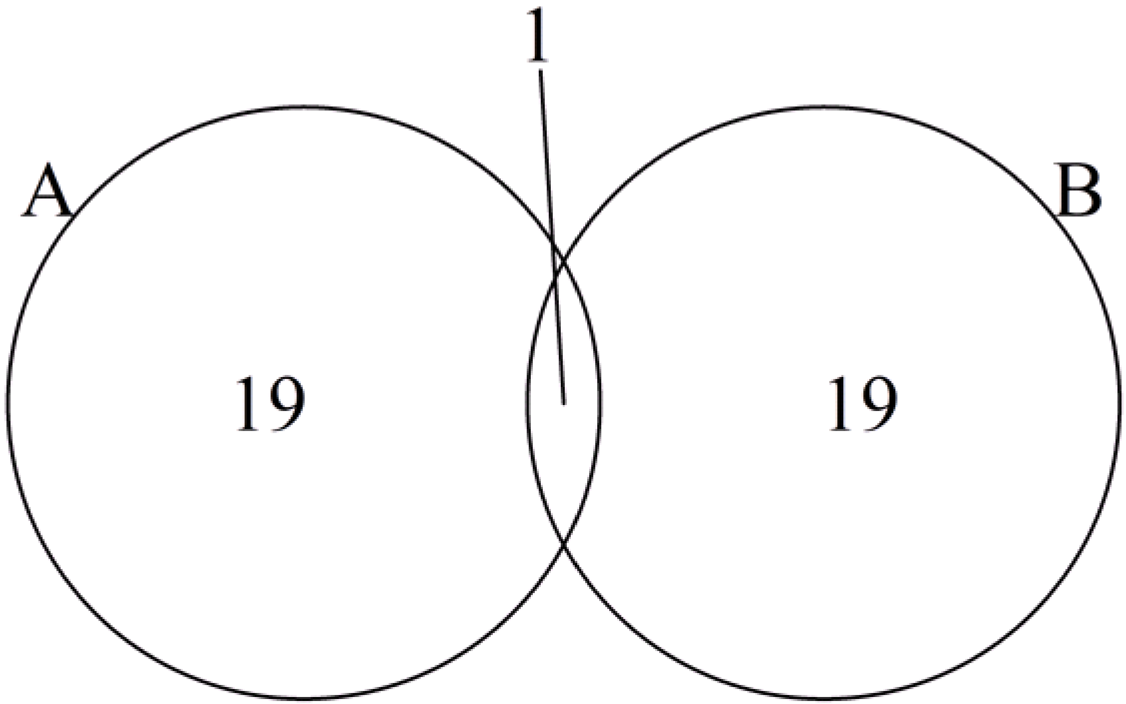
FIGURE B.13 VENN DIAGRAM STAGE3 TOP 20 RANKED GENES MOLENAAR AND WANG DATASETS



A = MOLENAAR STAGE 3 GRN TOP 20 RANKED GENES FOR COMBINED METRIC

B = WANG STAGE 3 GRN TOP 20 RANKED GENES FOR COMBINED METRIC

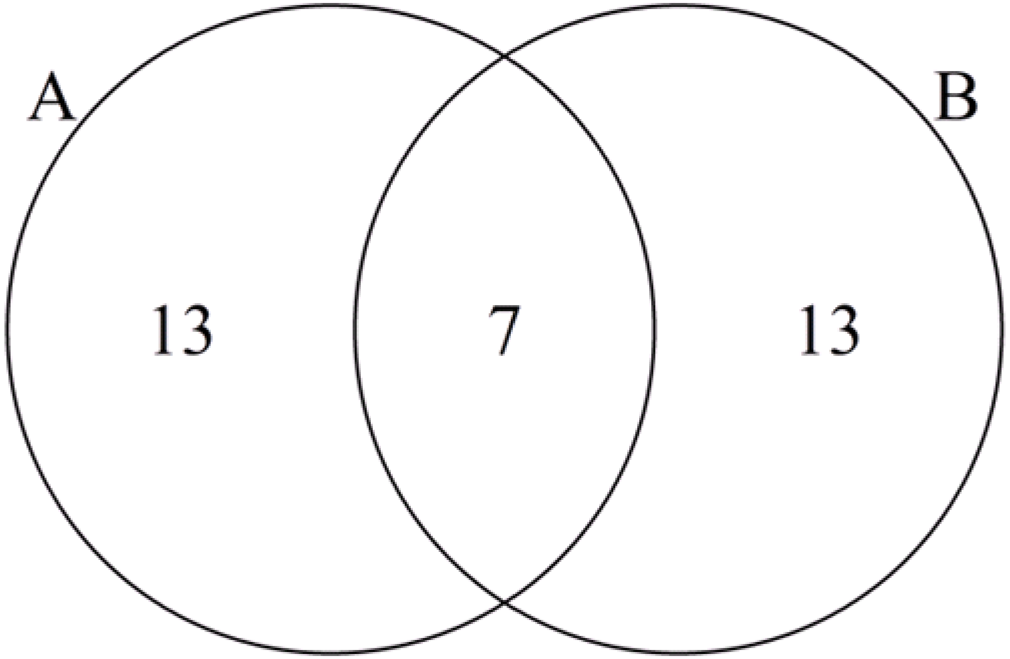
FIGURE B.14 VENN DIAGRAM STAGE4 TOP 20 RANKED GENES MOLENAAR AND WANG DATASETS



A = MOLENAAR STAGE 4 GRN TOP 20 RANKED GENES FOR COMBINED METRIC

B = WANG STAGE 4 GRN TOP 20 RANKED GENES FOR COMBINED METRIC

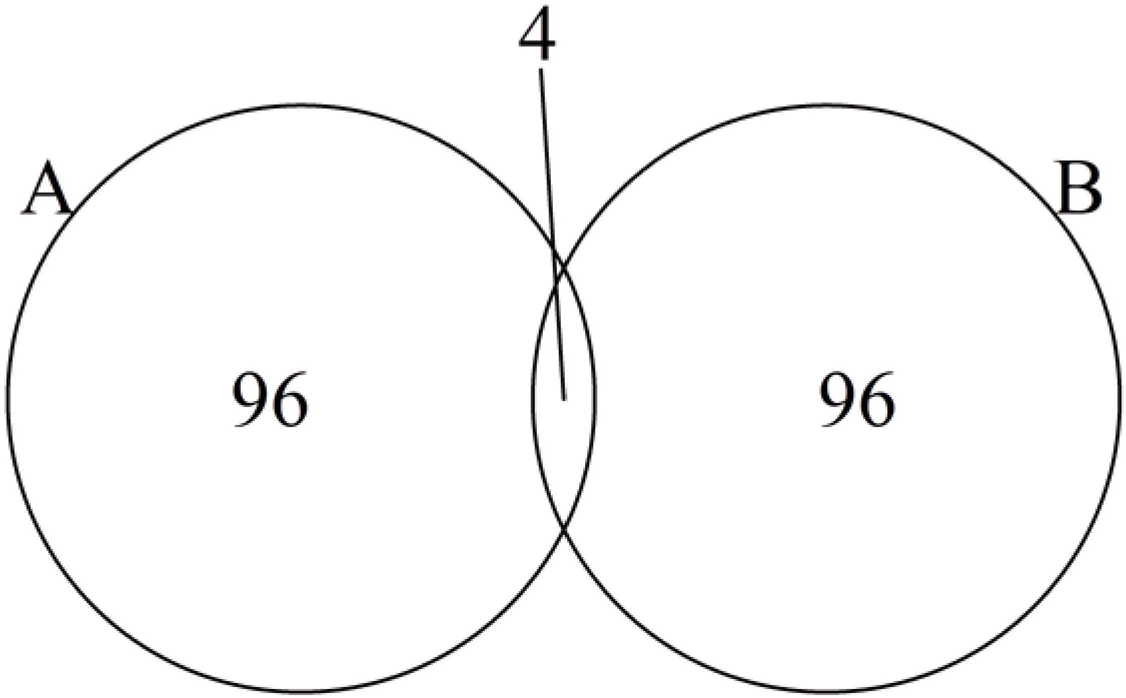
FIGURE B.15 VENN DIAGRAM STAGE4M TOP 20 RANKED GENES MOLENAAR AND WANG DATASETS



A = MOLENAAR STAGE 4M GRN TOP 20 RANKED GENES FOR COMBINED METRIC

B = WANG STAGE 4M GRN TOP 20 RANKED GENES FOR COMBINED METRIC

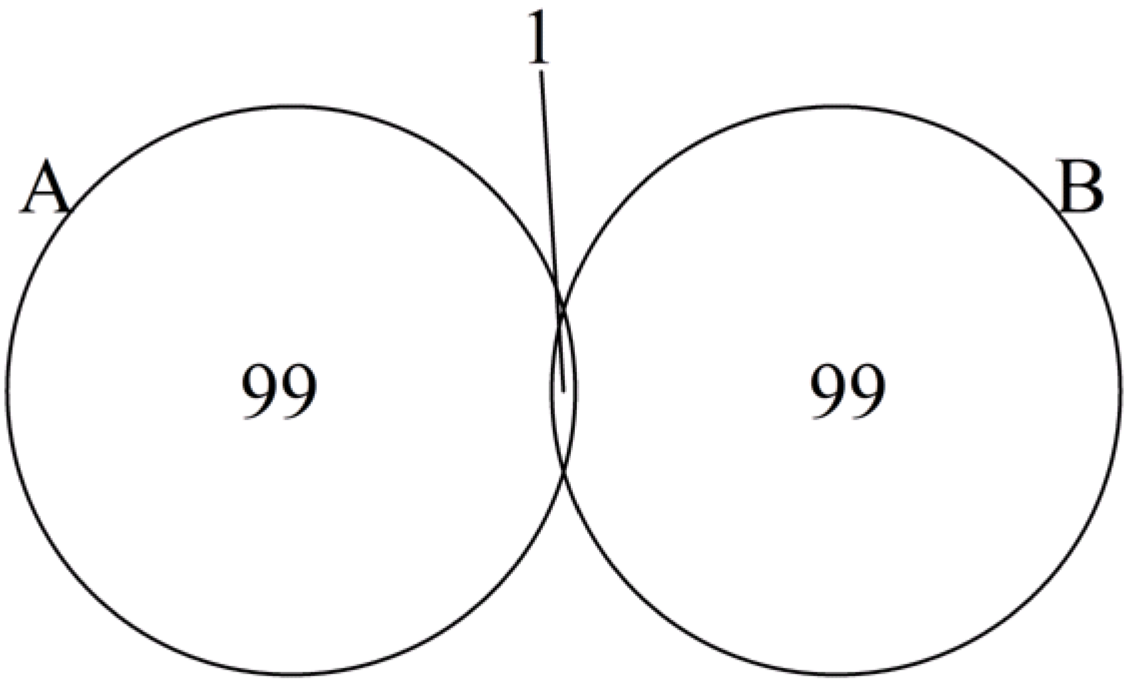
FIGURE B.16 VENN DIAGRAM STAGE1 TOP 100 RANKED GENES MOLENAAR AND WANG DATASETS



A = MOLENAAR STAGE 1 GRN TOP 100 RANKED GENES FOR COMBINED METRIC

B = WANG STAGE 1 GRN TOP 100 RANKED GENES FOR COMBINED METRIC

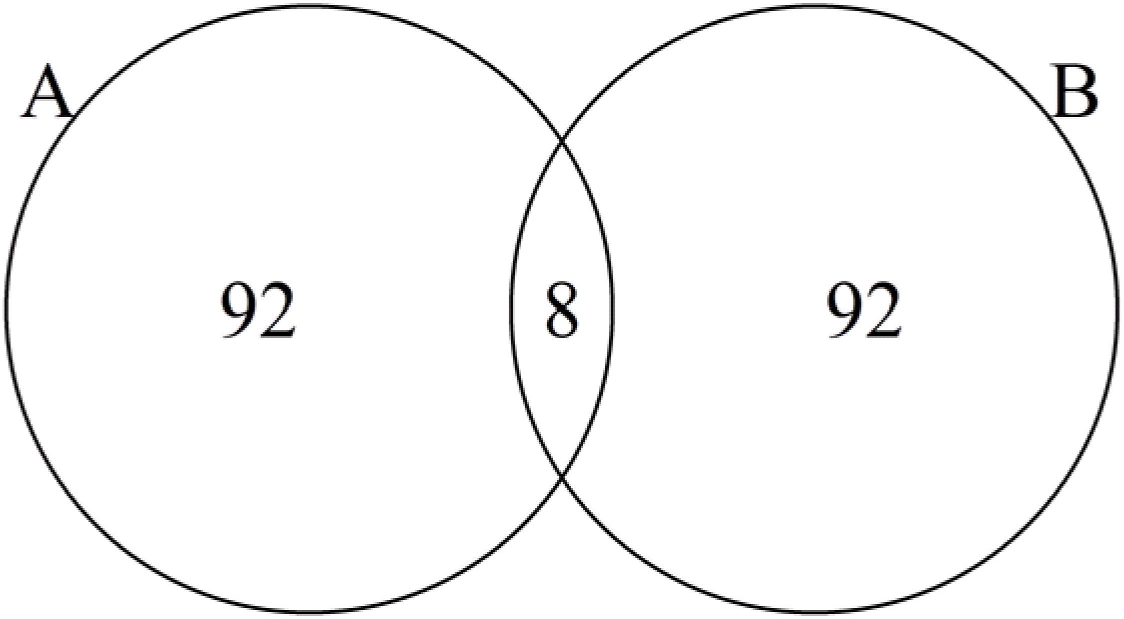
FIGURE B.17 VENN DIAGRAM STAGE3 TOP 100 RANKED GENES MOLENAAR AND WANG DATASETS



A = MOLENAAR STAGE 3 GRN TOP 100 RANKED GENES FOR COMBINED METRIC

B = WANG STAGE 3 GRN TOP 100 RANKED GENES FOR COMBINED METRIC

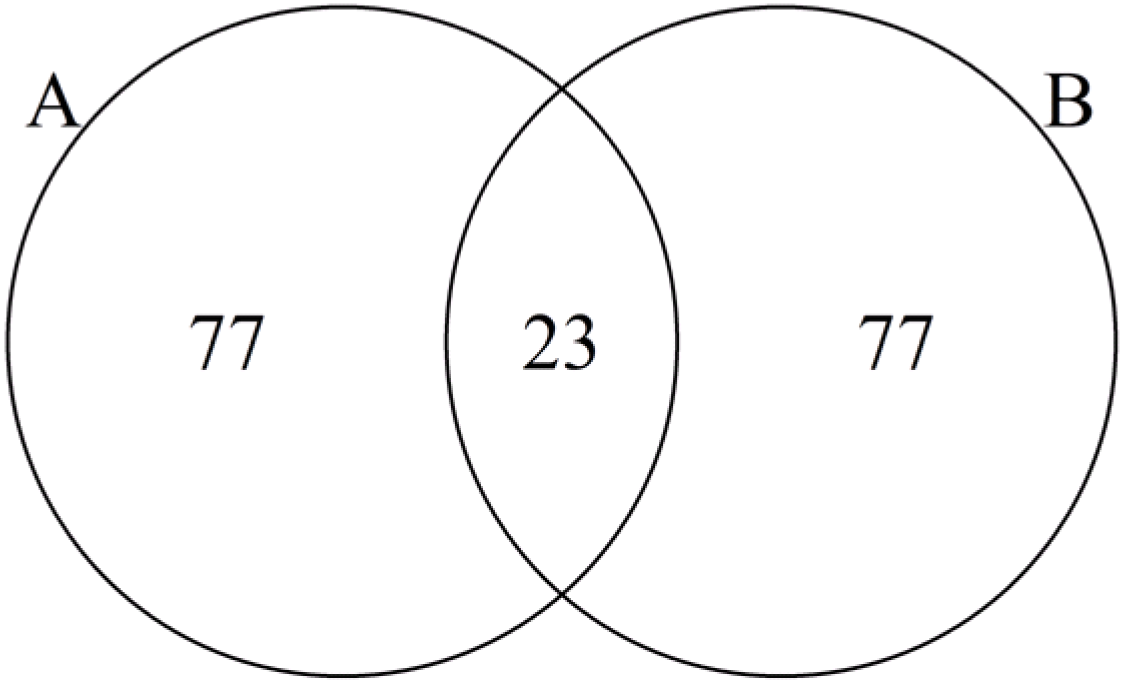
FIGURE B.18 VENN DIAGRAM STAGE4 TOP 100 RANKED GENES MOLENAAR AND WANG DATASETS



A = MOLENAAR STAGE 4 GRN TOP 100 RANKED GENES FOR COMBINED METRIC

B = WANG STAGE 4 GRN TOP 100 RANKED GENES FOR COMBINED METRIC

FIGURE B.19 VENN DIAGRAM STAGE4M TOP 100 RANKED GENES MOLENAAR AND WANG DATASETS



A = MOLENAAR STAGE 4M GRN TOP 100 RANKED GENES FOR COMBINED METRIC

B = WANG STAGE 4M GRN TOP 100 RANKED GENES FOR COMBINED METRIC

FIGURE B.20 VENN DIAGRAM OF GENE DISTRIBUTION IN THE LARGEST UNIQUE CLIQUES IN THE GLIOBLASTOMA GRNS INFERRED USING THE POSITIVE Z SCORE METHOD

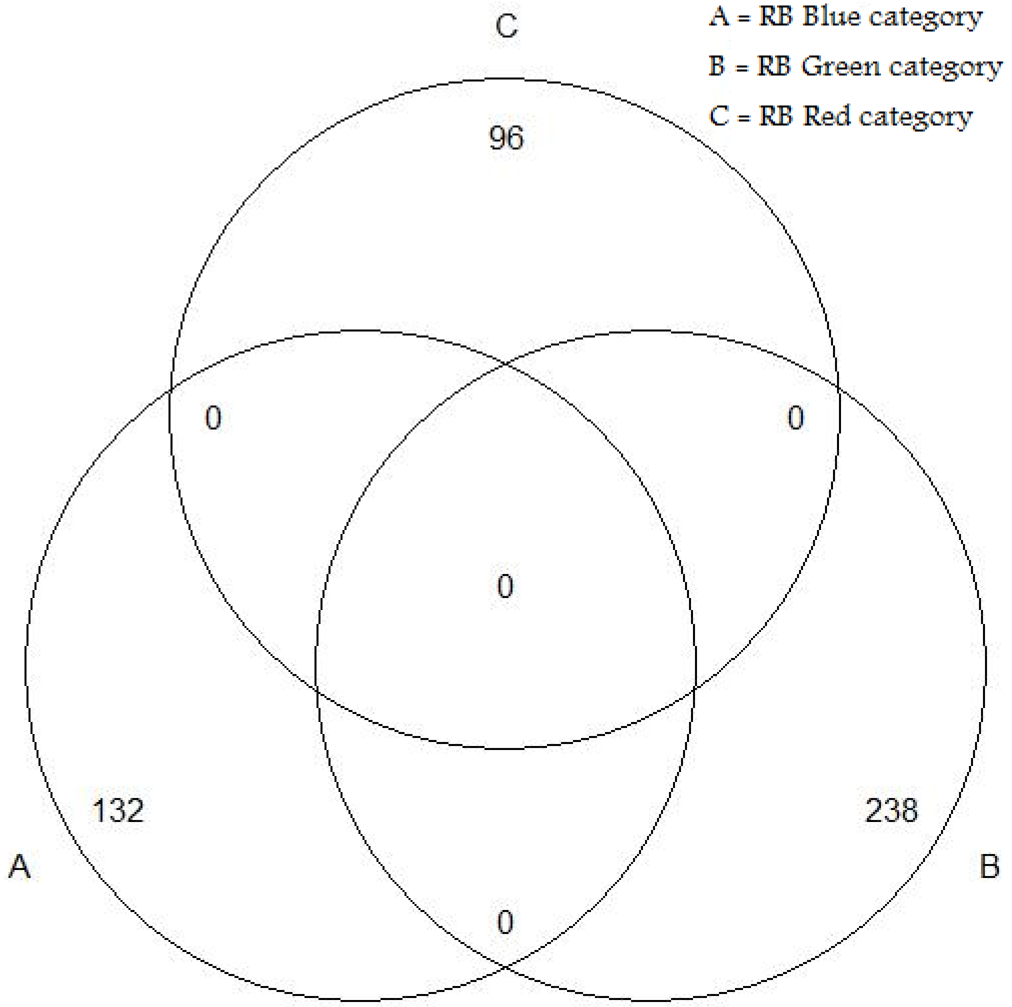


FIGURE B.21 GENE DISTRIBUTION OF TOP 20 COMBINED METRIC RANKINGS IN THE DIFFERENT CATEGORIES OF RETINOBLASTOMA GRN INFERRED USING THE POSITIVE Z SCORE METHOD

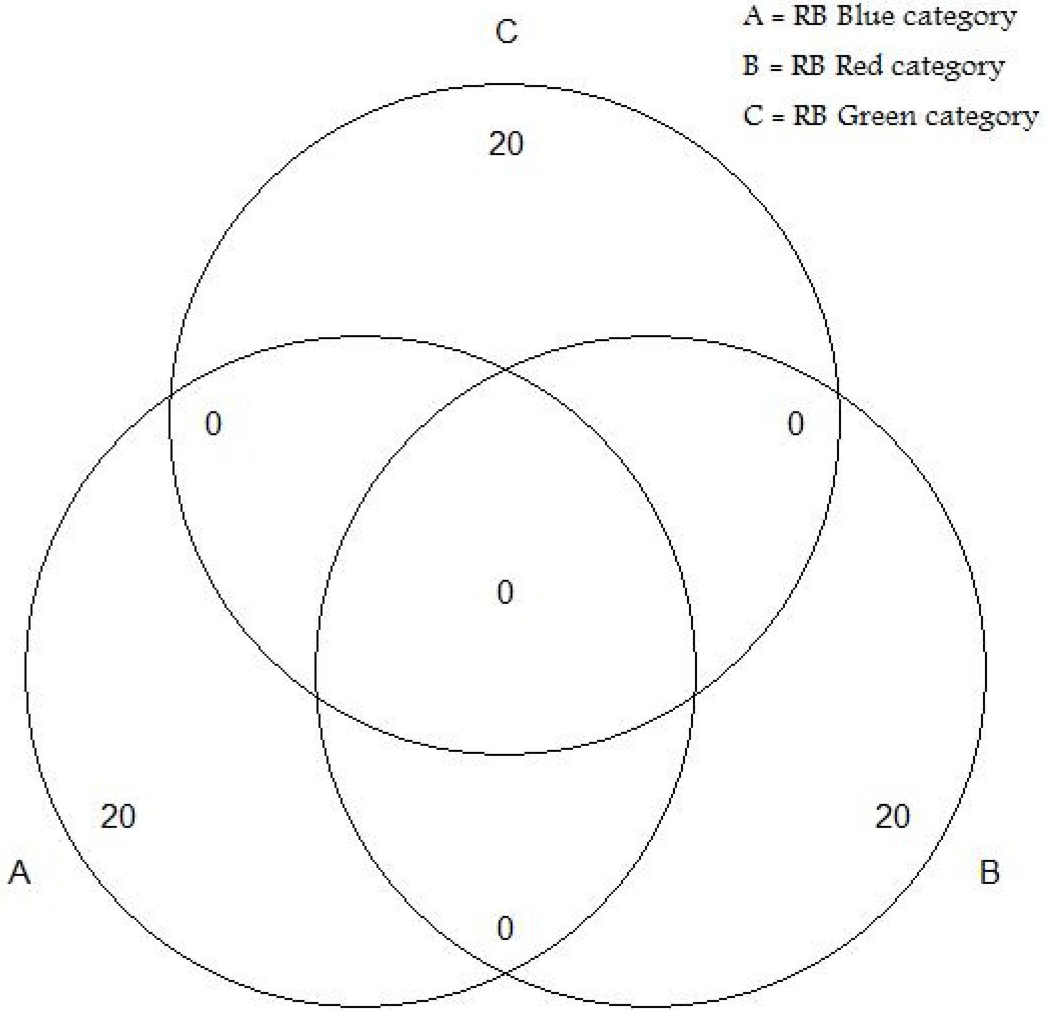


FIGURE B.22 GENE DISTRIBUTION OF TOP 100 COMBINED METRIC RANKINGS IN THE DIFFERENT CATEGORIES OF RETINOBLASTOMA GRN INFERRED USING THE POSITIVE Z SCORE METHOD

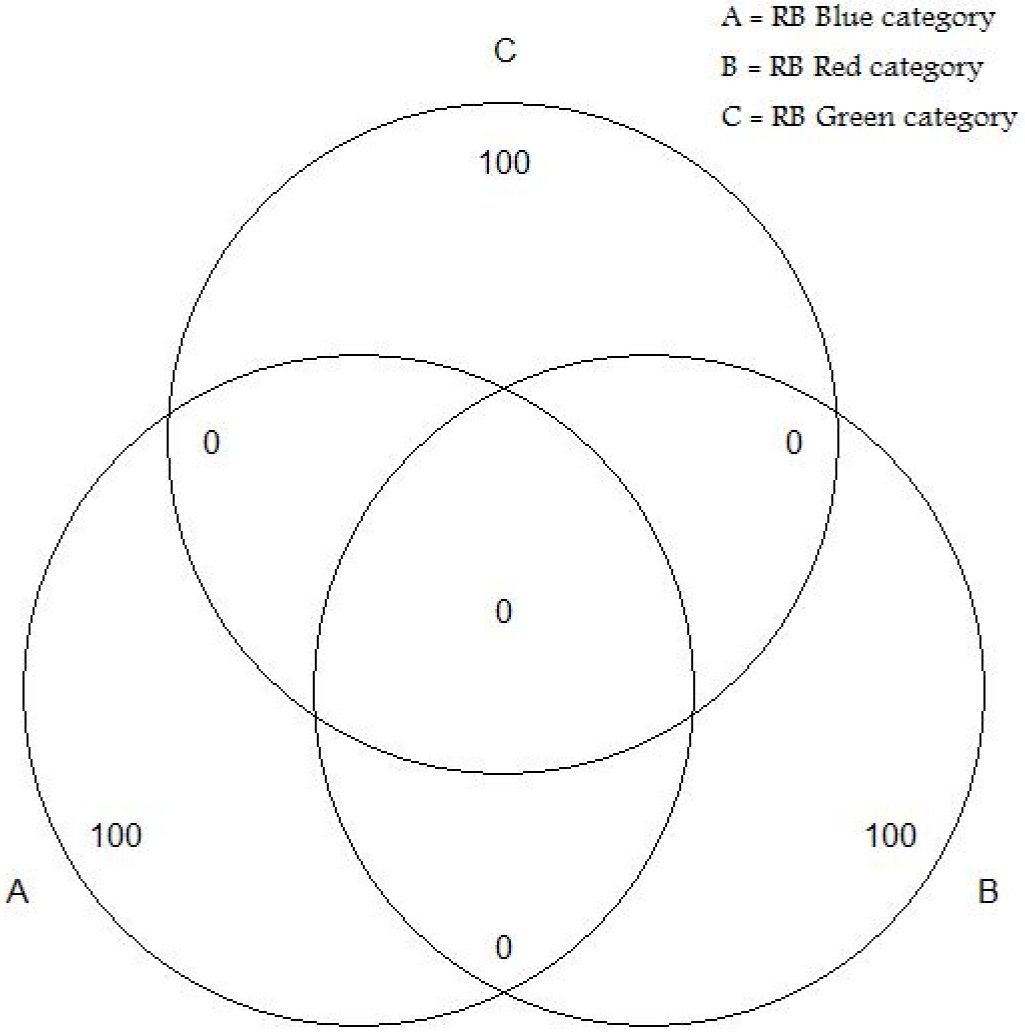
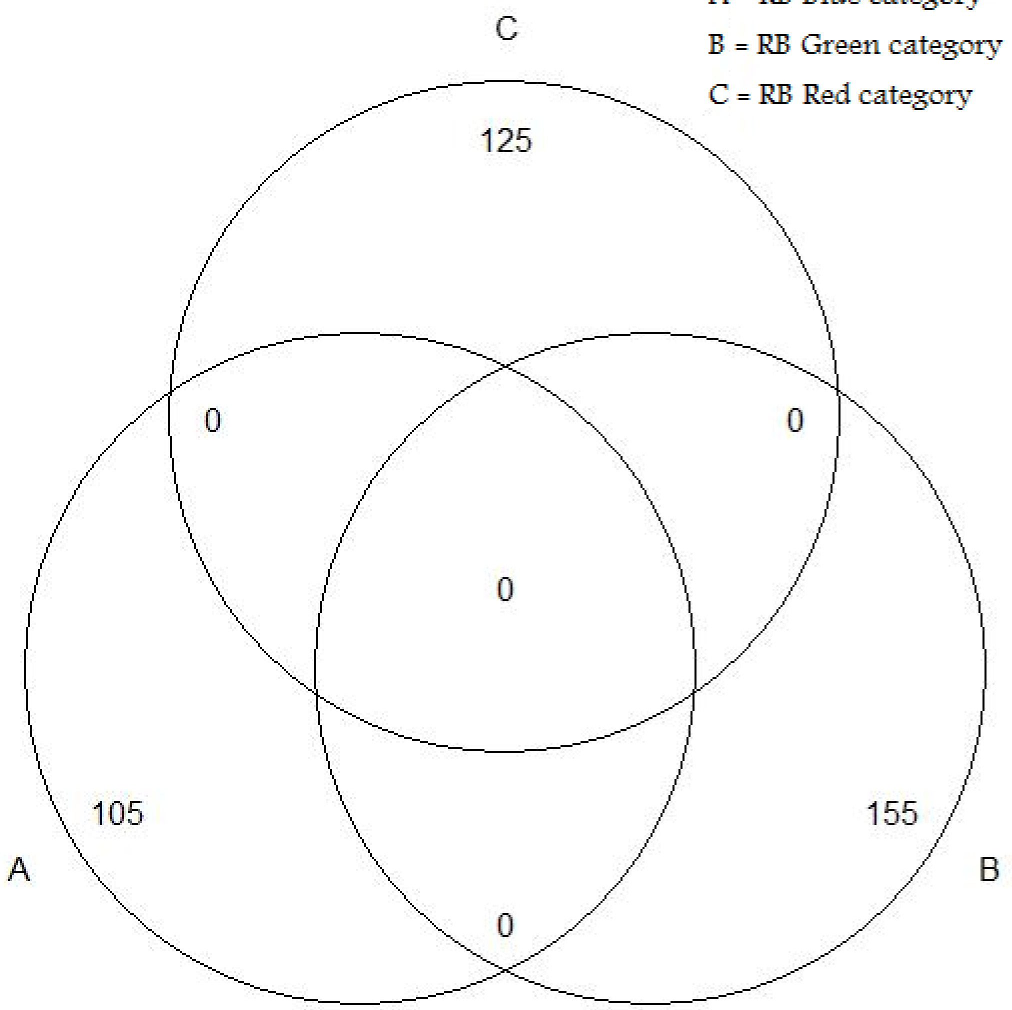


FIGURE B.23 VENN DIAGRAM OF GENE DISTRIBUTION IN THE LARGEST UNIQUE CLIQUES IN THE RETINOBLASTOMA GRNS INFERRED USING THE NEGATIVE Z SCORE METHOD



A = RB Blue category

B = RB Green category

C = RB Red category

FIGURE B.24 GENE DISTRIBUTION OF TOP 20 COMBINED METRIC RANKINGS IN THE DIFFERENT CATEGORIES OF RETINOBLASTOMA GRN INFERRED USING THE NEGATIVE Z SCORE METHOD

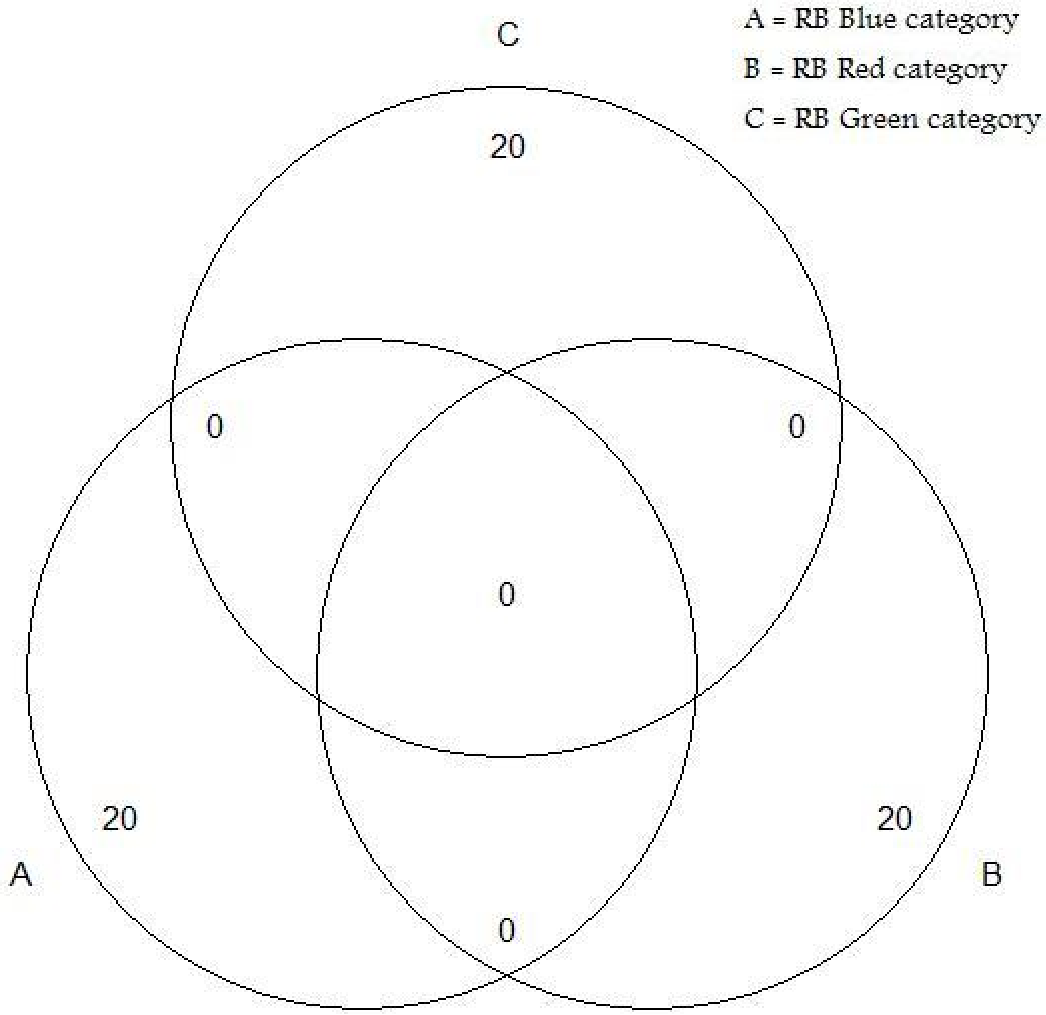
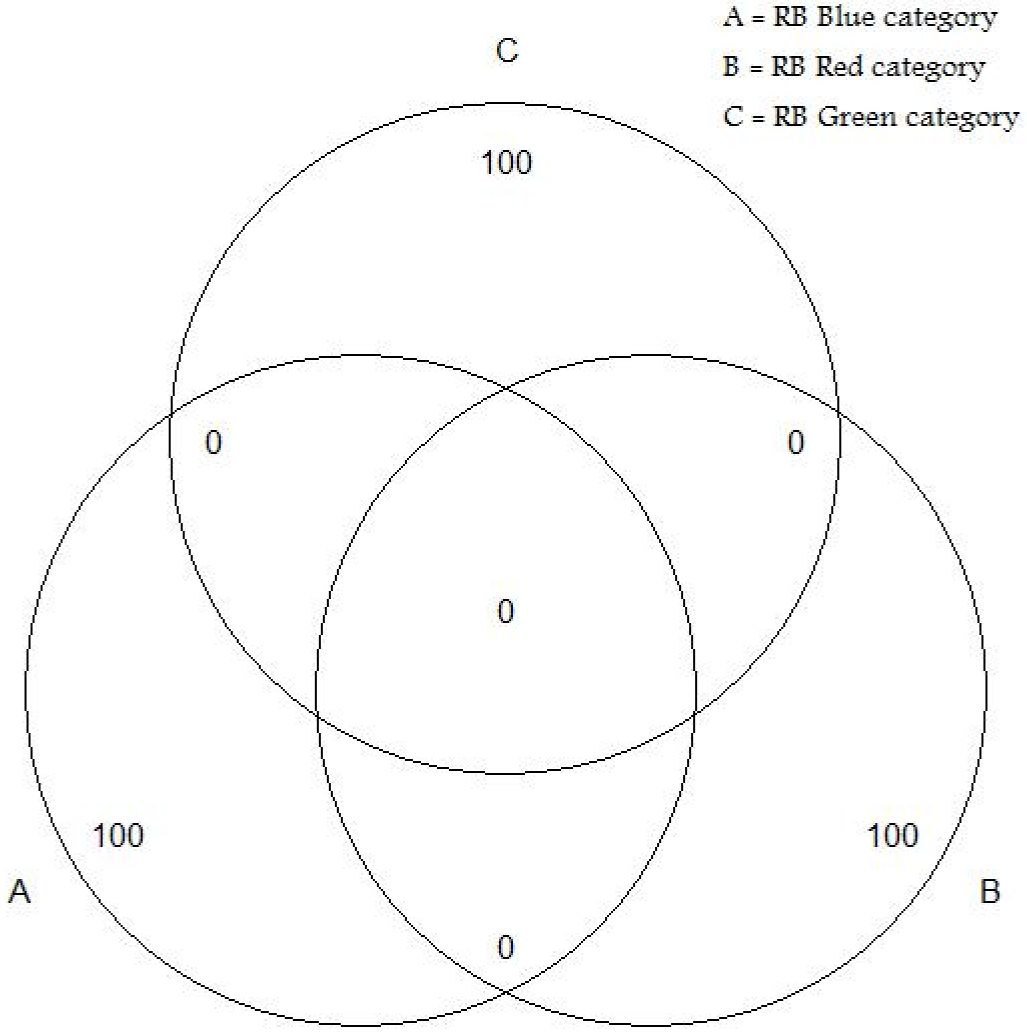


FIGURE B.25 GENE DISTRIBUTION OF TOP 100 COMBINED METRIC RANKINGS IN THE DIFFERENT CATEGORIES OF RETINOBLASTOMA GRN INFERRED USING THE NEGATIVE Z SCORE METHOD



A = RB Blue category

B = RB Red category

C = RB Green category

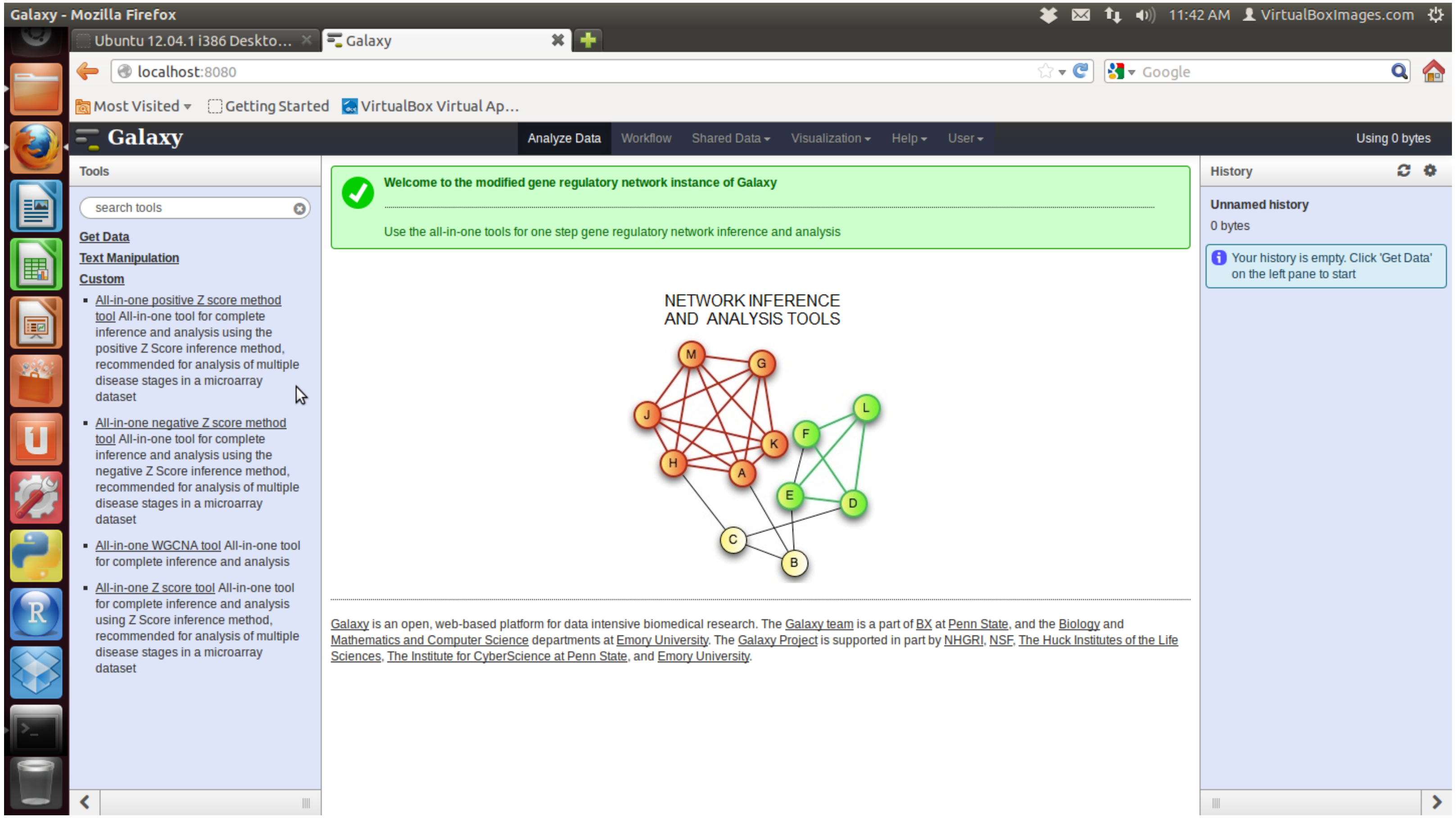
APPENDIX C

SCREENSHOTS OF GENE REGULATORY NETWORK INFERENCE AND ANALYSIS TOOLS ON GALAXY LOCAL HOST

The following appendix contains additional screenshots of various GRN inference and analysis tools implemented on a local host version of Galaxy introduced in chapter 8 of the work. One screenshot of both the all-in-one tool using the novel Z score GRN inference method, and the all-in-one tool using the WGCNA GRN inference method are provided. As the end user is presented with the same options for the all-in-one tools using the positive and negative Z score methods, these screenshots are not shown in order to avoid repetition.

A number of output files are provided to the end user; an adjacency matrix of the GRN inferred, a table with all the metric ranks for the genes in the network, the correlation scores between the rankings of the metrics calculated, enrichment results for the top ten ranked genes in the GRN based on the combined metric, and additionally edge lists of the whole network and the subnets of the top ten ranking genes that can be imported straight into Cytoscape for visualisation.

FIGURE C.1 SCREENSHOT OF TOOLS AVAILABLE TO THE END USER IN THE LOCAL HOST VERSION OF GALAXY



Welcome to the modified gene regulatory network instance of Galaxy

Use the all-in-one tools for one step gene regulatory network inference and analysis

- All-in-one positive Z score method tool All-in-one tool for complete inference and analysis using the positive Z Score inference method, recommended for analysis of multiple disease stages in a microarray dataset
- All-in-one negative Z score method tool All-in-one tool for complete inference and analysis using the negative Z Score inference method, recommended for analysis of multiple disease stages in a microarray dataset
- All-in-one WGCNA tool All-in-one tool for complete inference and analysis
- All-in-one Z score tool All-in-one tool for complete inference and analysis using Z Score inference method, recommended for analysis of multiple disease stages in a microarray dataset

Unnamed history

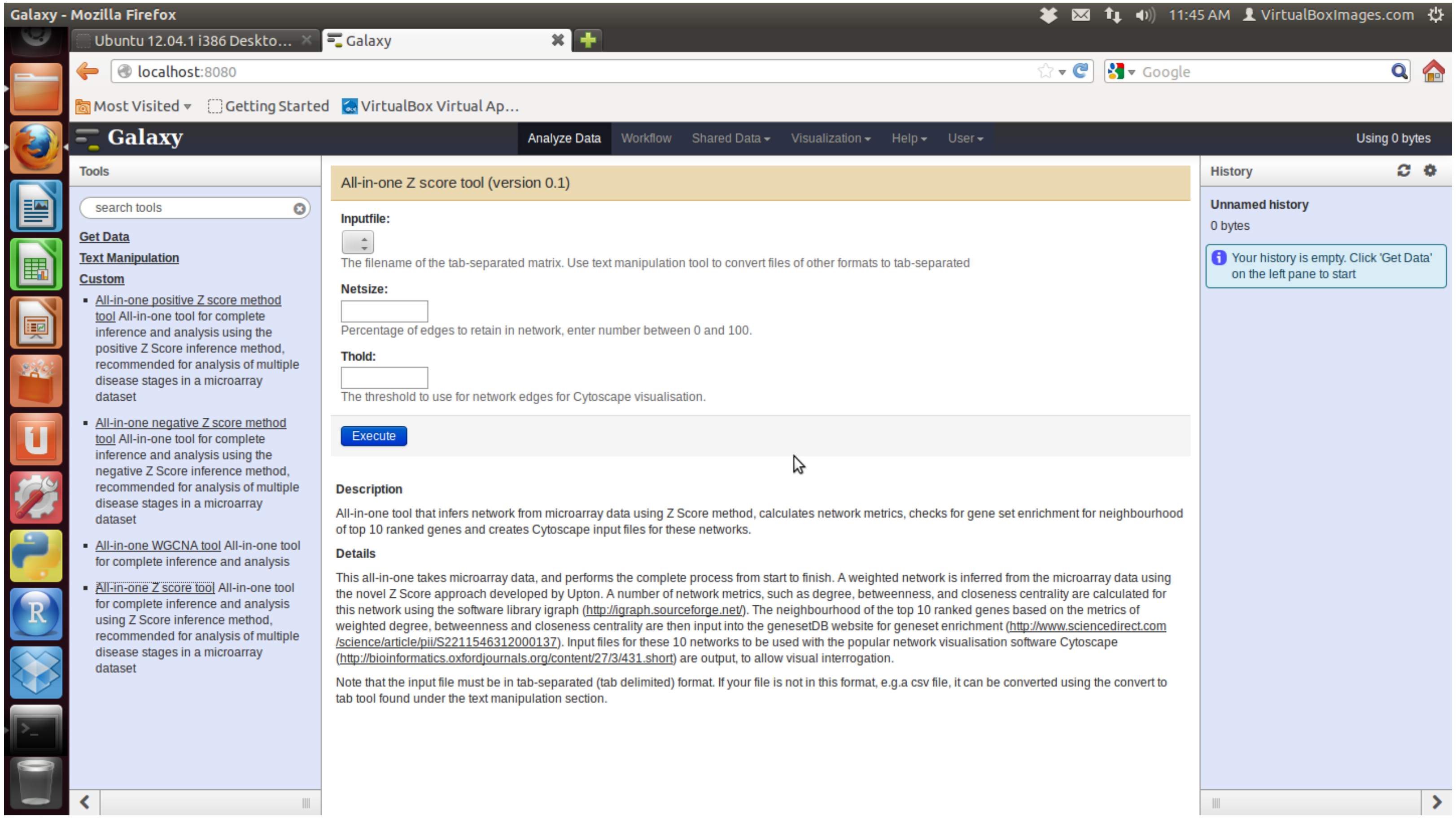
0 bytes

Your history is empty. Click 'Get Data' on the left pane to start

Galaxy is an open, web-based platform for data intensive biomedical research. The Galaxy team is a part of BX at Penn State, and the Biology and Mathematics and Computer Science departments at Emory University. The Galaxy Project is supported in part by NHGRI, NSF, The Huck Institutes of the Life Sciences, The Institute for CyberScience at Penn State, and Emory University.

As can be seen from the screenshot above, there are four all-in-one tools available to the end user. These are shown in more detail in the following screenshots.

FIGURE C.2 SCREENSHOT OF ALL-IN-ONE TOOL USING THE NOVEL Z SCORE GRN INFERENCE METHOD



Get Data

Text Manipulation

Custom

- All-in-one positive Z score method tool All-in-one tool for complete inference and analysis using the positive Z Score inference method, recommended for analysis of multiple disease stages in a microarray dataset
- All-in-one negative Z score method tool All-in-one tool for complete inference and analysis using the negative Z Score inference method, recommended for analysis of multiple disease stages in a microarray dataset
- All-in-one WGCNA tool All-in-one tool for complete inference and analysis
- All-in-one Z score tool All-in-one tool for complete inference and analysis using Z Score inference method, recommended for analysis of multiple disease stages in a microarray dataset

All-in-one Z score tool (version 0.1)

Inputfile:

The filename of the tab-separated matrix. Use text manipulation tool to convert files of other formats to tab-separated

Netsize:

Percentage of edges to retain in network, enter number between 0 and 100.

Thold:

The threshold to use for network edges for Cytoscape visualisation.

Description

All-in-one tool that infers network from microarray data using Z Score method, calculates network metrics, checks for gene set enrichment for neighbourhood of top 10 ranked genes and creates Cytoscape input files for these networks.

Details

This all-in-one takes microarray data, and performs the complete process from start to finish. A weighted network is inferred from the microarray data using the novel Z Score approach developed by Upton. A number of network metrics, such as degree, betweenness, and closeness centrality are calculated for this network using the software library igraph (http://igraph.sourceforge.net/). The neighbourhood of the top 10 ranked genes based on the metrics of weighted degree, betweenness and closeness centrality are then input into the genesetDB website for geneset enrichment (http://www.sciencedirect.com/science/article/pii/S2211546312000137). Input files for these 10 networks to be used with the popular network visualisation software Cytoscape (http://bioinformatics.oxfordjournals.org/content/27/3/431.short) are output, to allow visual interrogation.

Note that the input file must be in tab-separated (tab delimited) format. If your file is not in this format, e.g.a csv file, it can be converted using the convert to tab tool found under the text manipulation section.

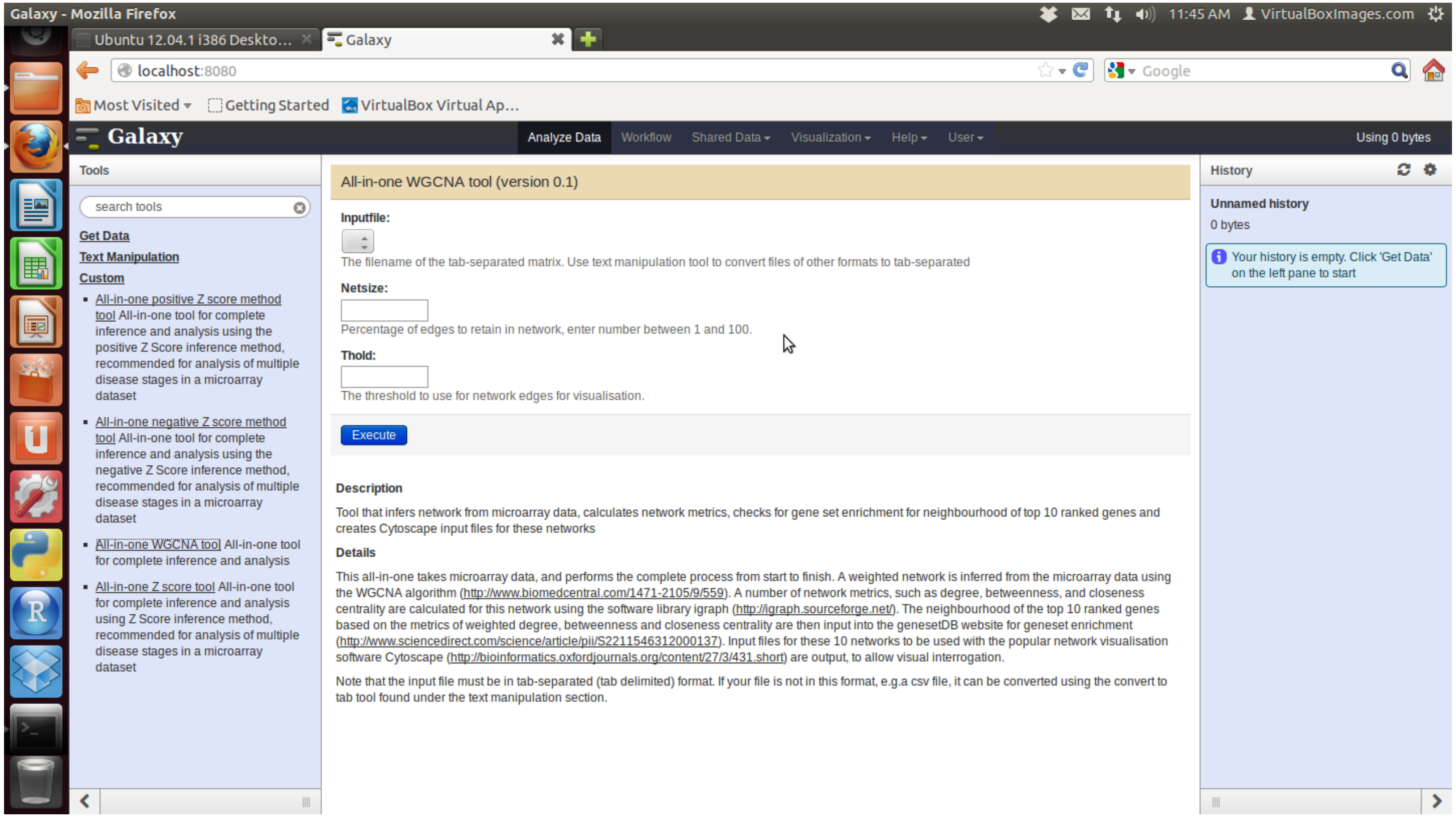
Unnamed history

0 bytes

Your history is empty. Click 'Get Data' on the left pane to start

The above screenshot is for the all-in-one tool using the novel Z score GRN inference method. The end user has three options; the input file location of the microarray dataset, the percentage of edges to retain in the network, and also the threshold of edge size to use for visualisation.

FIGURE C.3 SCREENSHOT OF ALL-IN-ONE TOOL USING THE WGCNA GRN INFERENCE METHOD



Get Data

Text Manipulation

Custom

- All-in-one positive Z score method tool All-in-one tool for complete inference and analysis using the positive Z Score inference method, recommended for analysis of multiple disease stages in a microarray dataset
- All-in-one negative Z score method tool All-in-one tool for complete inference and analysis using the negative Z Score inference method, recommended for analysis of multiple disease stages in a microarray dataset
- All-in-one WGCNA tool All-in-one tool for complete inference and analysis
- All-in-one Z score tool All-in-one tool for complete inference and analysis using Z Score inference method, recommended for analysis of multiple disease stages in a microarray dataset

All-in-one WGCNA tool (version 0.1)

Inputfile:

The filename of the tab-separated matrix. Use text manipulation tool to convert files of other formats to tab-separated

Netsize:

Percentage of edges to retain in network, enter number between 1 and 100.

Thold:

The threshold to use for network edges for visualisation.

Description

Tool that infers network from microarray data, calculates network metrics, checks for gene set enrichment for neighbourhood of top 10 ranked genes and creates Cytoscape input files for these networks

Details

This all-in-one takes microarray data, and performs the complete process from start to finish. A weighted network is inferred from the microarray data using the WGCNA algorithm (http://www.biomedcentral.com/1471-2105/9/559). A number of network metrics, such as degree, betweenness, and closeness centrality are calculated for this network using the software library igraph (http://igraph.sourceforge.net/). The neighbourhood of the top 10 ranked genes based on the metrics of weighted degree, betweenness and closeness centrality are then input into the genesetDB website for geneset enrichment (http://www.sciencedirect.com/science/article/pii/S2211546312000137). Input files for these 10 networks to be used with the popular network visualisation software Cytoscape (http://bioinformatics.oxfordjournals.org/content/27/3/431.short) are output, to allow visual interrogation.

Note that the input file must be in tab-separated (tab delimited) format. If your file is not in this format, e.g.a csv file, it can be converted using the convert to tab tool found under the text manipulation section.

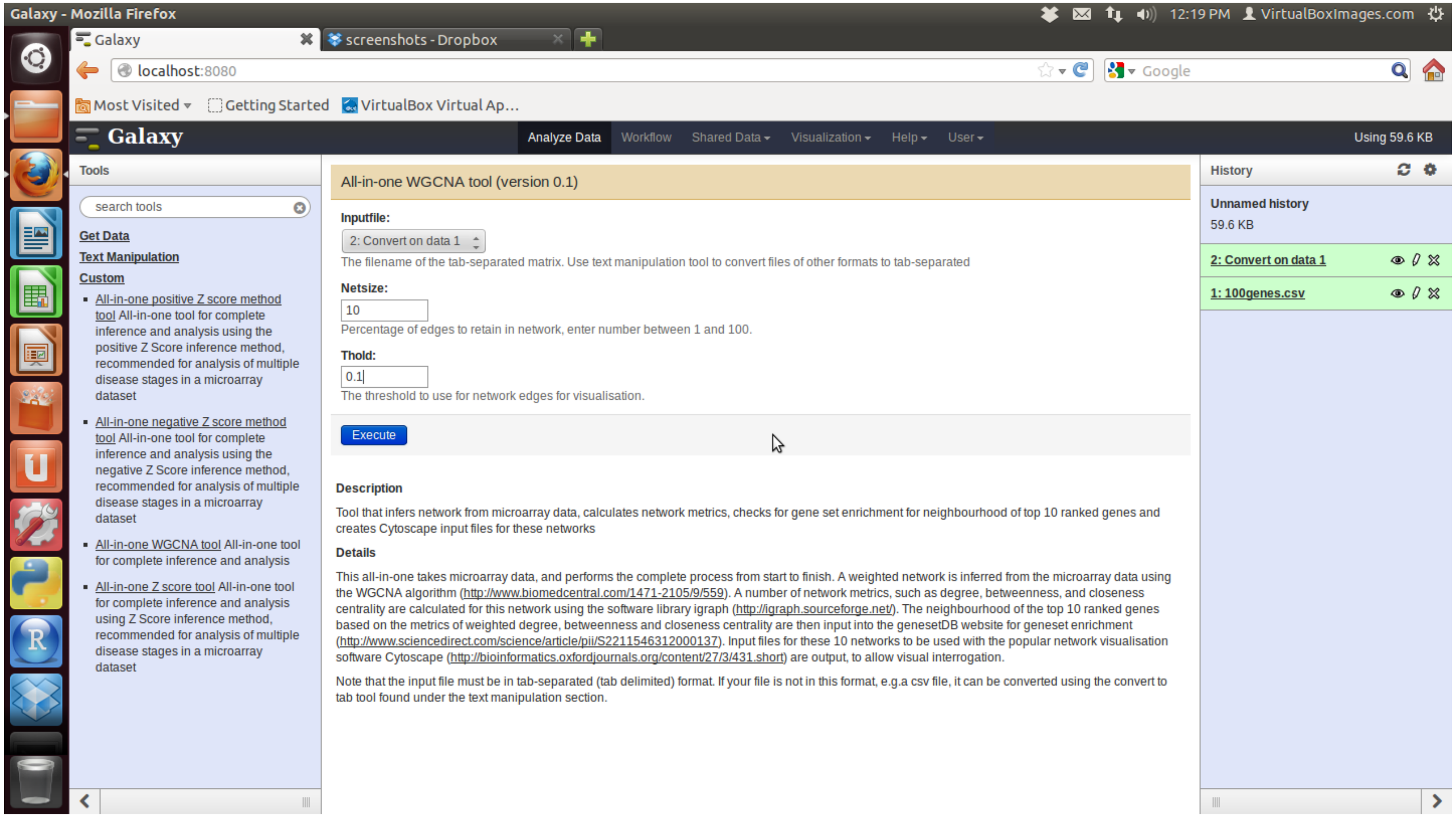
Unnamed history

0 bytes

Your history is empty. Click 'Get Data' on the left pane to start

The above screenshot is for the all-in-one tool using the WGCNA GRN inference method. The end user has the same three options; the input file location of the microarray dataset, the percentage of edges to retain in the network, and also the threshold of edge size to use for visualisation.

FIGURE C.4 SCREENSHOT OF ALL-IN-ONE TOOL USING THE WGCNA GRN INFERENCE METHOD



- All-in-one positive Z score method tool All-in-one tool for complete inference and analysis using the positive Z Score inference method, recommended for analysis of multiple disease stages in a microarray dataset
- All-in-one negative Z score method tool All-in-one tool for complete inference and analysis using the negative Z Score inference method, recommended for analysis of multiple disease stages in a microarray dataset
- All-in-one WGCNA tool All-in-one tool for complete inference and analysis
- All-in-one Z score tool All-in-one tool for complete inference and analysis using Z Score inference method, recommended for analysis of multiple disease stages in a microarray dataset

All-in-one WGCNA tool (version 0.1)

Inputfile:

2: Convert on data 1

The filename of the tab-separated matrix. Use text manipulation tool to convert files of other formats to tab-separated

Netsize:

10

Percentage of edges to retain in network, enter number between 1 and 100.

Thold:

0.1

The threshold to use for network edges for visualisation.

Description

Tool that infers network from microarray data, calculates network metrics, checks for gene set enrichment for neighbourhood of top 10 ranked genes and creates Cytoscape input files for these networks

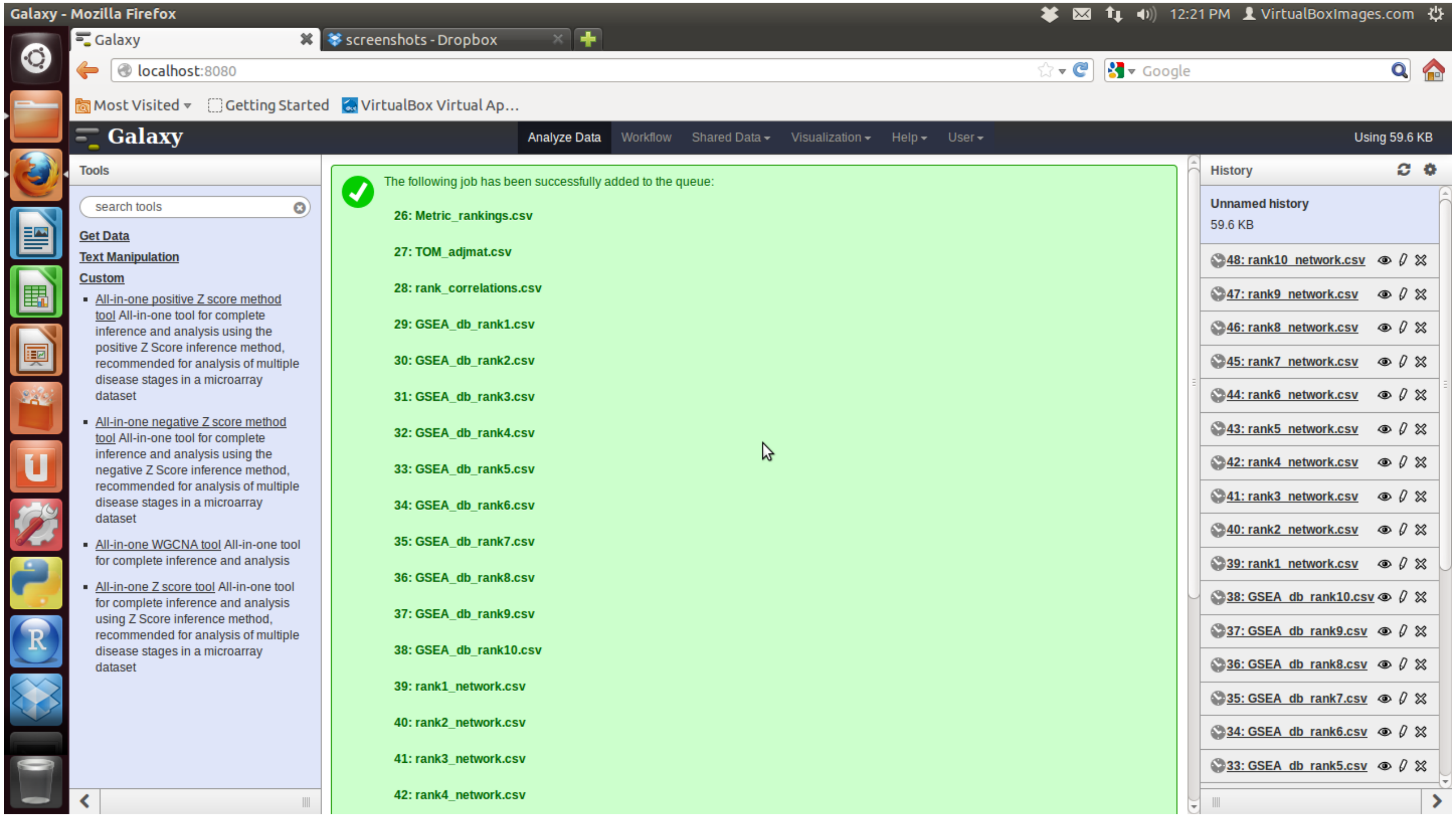
Details

This all-in-one takes microarray data, and performs the complete process from start to finish. A weighted network is inferred from the microarray data using the WGCNA algorithm (http://www.biomedcentral.com/1471-2105/9/559). A number of network metrics, such as degree, betweenness, and closeness centrality are calculated for this network using the software library igraph (http://igraph.sourceforge.net/). The neighbourhood of the top 10 ranked genes based on the metrics of weighted degree, betweenness and closeness centrality are then input into the genesetDB website for geneset enrichment (http://www.sciencedirect.com/science/article/pii/S2211546312000137). Input files for these 10 networks to be used with the popular network visualisation software Cytoscape (http://bioinformatics.oxfordjournals.org/content/27/3/431.short) are output, to allow visual interrogation.

Note that the input file must be in tab-separated (tab delimited) format. If your file is not in this format, e.g.a csv file, it can be converted using the convert to tab tool found under the text manipulation section.

The above screenshot shows input to the three options in all-one-tool using WGCNA. Finally, the screenshot below shows the output that is generated.

FIGURE C.5 SCREENSHOT OF OUTPUT OF ALL-IN-ONE TOOL USING THE WGCNA GRN INFERENCE METHOD



Get Data

Text Manipulation

Custom

- All-in-one positive Z score method tool All-in-one tool for complete inference and analysis using the positive Z Score inference method, recommended for analysis of multiple disease stages in a microarray dataset
- All-in-one negative Z score method tool All-in-one tool for complete inference and analysis using the negative Z Score inference method, recommended for analysis of multiple disease stages in a microarray dataset
- All-in-one WGCNA tool All-in-one tool for complete inference and analysis
- All-in-one Z score tool All-in-one tool for complete inference and analysis using Z Score inference method, recommended for analysis of multiple disease stages in a microarray dataset

The following job has been successfully added to the queue:

26: Metric_rankings.csv

27: TOM_adjmat.csv

28: rank_correlations.csv

29: GSEA_db_rank1.csv

30: GSEA_db_rank2.csv

31: GSEA_db_rank3.csv

32: GSEA_db_rank4.csv

33: GSEA_db_rank5.csv

34: GSEA_db_rank6.csv

35: GSEA_db_rank7.csv

36: GSEA_db_rank8.csv

37: GSEA_db_rank9.csv

38: GSEA_db_rank10.csv

39: rank1_network.csv

40: rank2_network.csv

41: rank3_network.csv

42: rank4_network.csv

Unnamed history
59.6 KB

48: rank10_network.csv
47: rank9_network.csv
46: rank8_network.csv
45: rank7_network.csv
44: rank6_network.csv
43: rank5_network.csv
42: rank4_network.csv
41: rank3_network.csv
40: rank2_network.csv
39: rank1_network.csv
38: GSEA_db_rank10.csv
37: GSEA_db_rank9.csv
36: GSEA_db_rank8.csv
35: GSEA_db_rank7.csv
34: GSEA_db_rank6.csv
33: GSEA_db_rank5.csv

As can be seen in the above screenshot, output files are generated to the end user without having to carry out any programming using specialist programming languages.

References

[1] U. Sauer, M. Heinemann, and N. Zamboni, "GENETICS: Getting Closer to the Whole Picture," *Science,* vol. 316, pp. 550-551, April 27, 2007.

[2] A. C. Ahn, M. Tewari, C.-S. Poon, and R. S. Phillips, "The Limits of Reductionism in Medicine: Could Systems Biology Offer an Alternative?," *PLoS Med,* vol. 3, p. e208, 2006.

[3] H. Kitano, "Systems Biology: A Brief Overview," *Science,* vol. 295, pp. 1662-1664, March 1, 2002.

[4] A. P. Heath and L. E. Kavraki, "Computational challenges in systems biology," *Computer Science Review,* vol. 3, pp. 1-17, 2009.

[5] M. Hecker, S. Lambeck, S. Toepfer, E. Van Someren, and R. Guthke, "Gene regulatory network inference: data integration in dynamic modelsâ€"a review," *Biosystems,* vol. 96, pp. 86-103, 2009.

[6] E. Bonnet, T. Michoel, and Y. Van de Peer, "Prediction of a gene regulatory network linked to prostate cancer from gene expression, microRNA and clinical data," *Bioinformatics,* vol. 26, pp. i638-i644, September 15, 2010.

[7] J. Jeong and D. Lee, "Inferring candidate regulatory networks in human breast cancer cells," *Bioinformatics,* vol. 2, pp. 26-29, 2007.

[8] S. Horvath, B. Zhang, M. Carlson, K. Lu, S. Zhu, R. Felciano, M. Laurance, W. Zhao, Q. Shu, Y. Lee, A. Scheck, L. Liau, H. Wu, D. Geschwind, P. Febbo, H. Kornblum, T. Cloughesy, S. Nelson, and P. Mischel, "Analysis of Oncogenic Signaling Networks in Glioblastoma Identifies ASPM as a Novel Molecular Target," *Proc Natl Acad Sci USA,* vol. 103, pp. 17402 - 17407, 2006.

[9] D. Hurley, H. Araki, Y. Tamada, B. Dunmore, D. Sanders, S. Humphreys, M. Affara, S. Imoto, K. Yasuda, Y. Tomiyasu, K. Tashiro, C. Savoie, V. Cho, S. Smith, S. Kuhara, S. Miyano, D. S. Charnock-Jones, E. J. Crampin, and C. G. Print, "Gene network inference and visualization tools for biologists: application to new human transcriptome datasets," *Nucleic Acids Research,* November 24, 2011.

[10] J. Goecks, A. Nekrutenko, J. Taylor, and T. The Galaxy, "Galaxy: a comprehensive approach for supporting accessible, reproducible, and transparent computational research in the life sciences," *Genome Biology,* vol. 11, p. R86, 2010.

[11] J.-F. Fontaine, F. Priller, A. Barbosa-Silva, and M. A. Andrade-Navarro, "Génie: literature-based gene prioritization at multi genomic scale," *Nucleic Acids Research,* vol. 39, pp. W455-W461, July 1, 2011.

[12] L. H. Hartwell, J. J. Hopfield, S. Leibler, and A. W. Murray, "From molecular to modular cell biology," *Nature,* 1999.

[13] L. Hood, "Systems biology: integrating technology, biology, and computation," *Mechanisms of Ageing and Development,* vol. 124, pp. 9-16, 2003.

[14] M. W. Covert, E. M. Knight, J. L. Reed, M. J. Herrgard, and B. O. Palsson, "Integrating high-throughput and computational data elucidates bacterial networks," *Nature,* vol. 429, pp. 92-96, 2004.

[15] E. Alm and A. P. Arkin, "Biological Networks," *Current Opinion in Structural Biology,* vol. 13, pp. 193-202, 2003.

[16] U. Stelzl, U. Worm, M. Lalowski, C. Haenig, F. H. Brembeck, H. Goehler, M. Stroedicke, M. Zenkner, A. Schoenherr, and S. Koeppen, "A human protein-protein interaction network: a resource for annotating the proteome," *Cell,* vol. 122, pp. 957-968, 2005.

[17] P. Uetz, L. Giot, G. Cagney, T. A. Mansfield, R. S. Judson, J. R. Knight, D. Lockshon, V. Narayan, M. Srinivasan, and P. Pochart, "A comprehensive analysis of protein-protein interactions in Saccharomyces cerevisiae," *Nature,* vol. 403, pp. 623-627, 2000.

[18] J.-F. Rual, K. Venkatesan, T. Hao, T. Hirozane-Kishikawa, A. Dricot, N. Li, G. F. Berriz, F. D. Gibbons, M. Dreze, and N. Ayivi-Guedehoussou, "Towards a proteome-scale map of the human protein-protein interaction network," *Nature,* vol. 437, pp. 1173-1178, 2005.

[19] L. Giot, J. S. Bader, C. Brouwer, A. Chaudhuri, B. Kuang, Y. Li, Y. L. Hao, C. E. Ooi, B. Godwin, and E. Vitols, "A protein interaction map of Drosophila melanogaster," *Science,* vol. 302, pp. 1727-1736, 2003.

[20] J. Xu and Y. Li, "Discovering disease-genes by topological features in human protein-protein interaction network," *Bioinformatics,* vol. 22, pp. 2800 - 2805, 2006.

[21] S. Klamt and J. Stelling, "Two approaches for metabolic pathway analysis?," *Trends in Biotechnology,* vol. 21, pp. 64-69, 2003.

[22] R. Steuer, "Computational approaches to the topology, stability and dynamics of metabolic networks," *Phytochemistry,* vol. 68, pp. 2139-2151, 2007.

[23] M. Kanehisa, "A database for post-genome analysis," *Trends in genetics: TIG,* vol. 13, p. 375, 1997.

[24] F. Vladimir, "Identifying Gene Regulatory Networks from Gene Expression Data," in *Handbook of Computational Molecular Biology*: Chapman and Hall/CRC, 2005, pp. 27-1-27-29.

[25] U. Mansmann and V. Jurinovic, "Biological feature validation of estimated gene interaction networks from microarray data: a case study on MYC in lymphomas," *Briefings in bioinformatics,* vol. 12, pp. 230-244.

[26] F. Jacob, "Evolution and tinkering," *Science,* vol. 196, pp. 1161-1166, 1977.

[27] U. Alon, "Biological Networks: The Tinkerer as an Engineer," *Science,* vol. 301, pp. 1866-1867, 2003.

[28] A.-L. Barabasi and Z. N. Oltvai, "Network biology: understanding the cell's functional organization," *Nat Rev Genet,* vol. 5, pp. 101-113, 2004.

[29] C. R. Myers, "Software systems as complex networks: Structure, function, and evolvability of software collaboration graphs," *Physical Review E,* vol. 68, p. 046116, 2003.

[30] U. Alon, M. G. Surette, N. Barkai, and S. Leibler, "Robustness in bacterial chemotaxis," *Nature,* vol. 397, pp. 168-171, 1999.

[31] A. Eldar, R. Dorfman, D. Weiss, H. Ashe, B.-Z. Shilo, and N. Barkai, "Robustness of the BMP morphogen gradient in Drosophila embryonic patterning," *Nature,* vol. 419, pp. 304-308, 2002.

[32] R. Milo, S. Shen-Orr, S. Itzkovitz, N. Kashtan, D. Chklovskii, and U. Alon, "Network motifs: simple building blocks of complex networks," *Science,* vol. 298, pp. 824 - 827, 2002.

[33] J. L. Gross and J. Yellen, *Graph theory and its applications*: Chapman & Hall/CRC, 2006.

[34] D. B. West, *Introduction to graph theory* vol. 2: Prentice hall Englewood Cliffs, 2001.

[35] M. E. J. Newman, "Analysis of weighted networks," *Physical Review E,* vol. 70, p. 056131, 2004.

[36] L. Euler, *The seven bridges of Konigsberg*: Wm. Benton, 1956.

[37] T. Zhou, G. Yan, and B.-H. Wang, "Maximal planar networks with large clustering coefficient and power-law degree distribution," *Physical Review E,* vol. 71, p. 046141, 2005.

[38] A. Rapoport, "Contribution to the theory of random and biased nets," *The bulletin of mathematical biophysics,* vol. 19, pp. 257-277, 1957.

[39] P. Erdős and A. Rényi, "On the evolution of random graphs," *Publ. Math. Inst. Hungar. Acad. Sci,* vol. 5, pp. 17-61, 1960.

[40] M. Girvan and M. E. J. Newman, "Community structure in social and biological networks," *Proceedings of the National Academy of Sciences of the United States of America,* vol. 99, pp. 7821-7826, 2002.

[41] D. J. Watts and S. H. Strogatz, "Collective dynamics of 'small-world' networks," *Nature,* vol. 393, pp. 440 - 442, 1998.

[42] S. Milgram, "The small world problem," *Psychology Today,* 1967.

[43] J. W. Grossman, *Erdos Number Project*: Jerrold W. Grossman., 2002.

[44] A.-L. Barabási and R. Albert, "Emergence of Scaling in Random Networks," *Science,* vol. 286, pp. 509-512, October 15, 1999.

[45] O. Hein, M. Schwind, and W. König, "Scale-free networks," *WIRTSCHAFTSINFORMATIK,* vol. 48, pp. 267-275, 2006.

[46] H. Jeong, B. Tombor, R. Albert, Z. N. Oltvai, and A. L. Barabási, "The large-scale organization of metabolic networks," *Nature,* vol. 407, pp. 651-654, 2000.

[47] J. Podani, Z. N. Oltvai, H. Jeong, B. Tombor, A. L. Barabási, and E. Szathmary, "Comparable system-level organization of Archaea and Eukaryotes," *Nat Genet,* vol. 29, pp. 54-56, 2001.

[48] B. Zhang and S. Horvath, "A General Framework for Weighted Gene Co-expression Network Analysis," *Stat Appl Genet Mol Biol,* vol. 4, p. Article 17, 2005.

[49] R. Cohen, K. Erez, D. ben-Avraham, and S. Havlin, "Resilience of the Internet to Random Breakdowns," *Physical Review Letters,* vol. 85, p. 4626, 2000.

[50] R. V. Solé, "Complex networks: structure, robustness and function," 2012.

[51] H. De Jong, "Modeling and Simulation of Genetic Regulatory Systems: A Literature Review," *Journal of Computational Biology,* vol. 9, pp. 67 - 103, 2002.

[52] Newman, "A measure of betweenness centrality based on random walks," *Social Networks,* vol. 27, pp. 39-54, 2005.

[53] T. Opsahl, F. Agneessens, and J. Skvoretz, "Node centrality in weighted networks: Generalizing degree and shortest paths," *Social Networks,* vol. 32, pp. 245-251, 2010.

[54] D. Koschützki and F. Schreiber, "Centrality analysis methods for biological networks and their application to gene regulatory networks," *Gene regulation and systems biology,* vol. 2, p. 193, 2008.

[55] L. Freeman, "Centrality in social networks: Conceptual clarification," *Social Networks,* vol. 1, pp. 215 - 239, 1978.

[56] G. Scardoni and C. Laudanna, "Centralities based analysis of complex networks," *New Frontiers in Graph Theory. InTech Open,* 2012.

[57] A. P. Potapov, N. Voss, N. Sasse, and E. Wingender, "Topology of mammalian transcription networks," *Genome Informatics Series,* vol. 16, p. 270, 2005.

[58] E. W. Dijkstra, "A note on two problems in connexion with graphs," *Numerische mathematik,* vol. 1, pp. 269-271, 1959.

[59] M. E. J. Newman, "Scientific collaboration networks. II. Shortest paths, weighted networks, and centrality," *Physical Review E,* vol. 64, p. 016132, 2001.

[60] U. Brandes, "A faster algorithm for betweenness centrality*," *Journal of Mathematical Sociology,* vol. 25, pp. 163-177, 2001.

[61] P. Bonacich, "Factoring and weighting approaches to status scores and clique identification," *The Journal of Mathematical Sociology,* vol. 2, pp. 113-120, 2013/04/04 1972.

[62] P. Bonacich, "Power and centrality: A family of measures," *American journal of sociology,* pp. 1170-1182, 1987.

[63] M. E. J. Newman, "Assortative Mixing in Networks," *Physical Review Letters,* vol. 89, p. 208701, 2002.

[64] M. W. Hahn, G. C. Conant, and A. Wagner, "Molecular evolution in large genetic networks: does connectivity equal constraint?," *Journal of Molecular Evolution,* vol. 58, pp. 203-211, 2004.

[65] Q.-J. Jiao and H.-B. Shen, "Maximum-clique algorithm: An effective method to mine large-scale co-expressed genes in Arabidopsis anther," in *2011 30th Chinese Control Conference (CCC)*, 2011, pp. 5650-5655.

[66] C. A. Ball and G. Sherlock, "What Are Microarrays?," *Bioinformatics for Geneticists: A Bioinformatics Primer for the Analysis of Genetic Data,* p. 371, 2007.

[67] R. A. Irizarry, B. Hobbs, F. Collin, Y. D. Beazer-Barclay, K. J. Antonellis, U. Scherf, and T. P. Speed, "Exploration, normalization and summaries of high density oligonucleotide array probe level data," *Biostatistics,* vol. 4, pp. 249-294, 2003.

[68] B. M. Bolstad, R. A. Irizarry, M. Astrand, and T. P. Speed, "A comparison of normalization methods for high density oligonucleotide array data based on variance and bias," *Bioinformatics,* vol. 19, pp. 185 - 193, 2003.

[69] J. M. Freudenberg, "Comparison of background correction and normalization procedures for high-density oligonucleotide microarrays," *Institut fur Informatik,* 2005.

[70] A. A. Hill, E. L. Brown, M. Z. Whitley, G. Tucker-Kellogg, C. P. Hunter, and D. K. Slonim, "Evaluation of normalization procedures for oligonucleotide array data based on spiked cRNA controls," *Genome Biol,* vol. 2, pp. 1-0055.13, 2001.

[71] P. Meyer, K. Kontos, F. Lafitte, and G. Bontempi, "Information-Theoretic Inference of Large Transcriptional Regulatory Networks," *EURASIP Journal on Bioinformatics and Systems Biology,* vol. 2007, p. 79879, 2007.

[72] P. Meyer, F. Lafitte, and G. Bontempi, "minet: A R/Bioconductor Package for Inferring Large Transcriptional Networks Using Mutual Information," *BMC Bioinformatics,* vol. 9, p. 461, 2008.

[73] A. Margolin, I. Nemenman, K. Basso, C. Wiggins, G. Stolovitzky, R. Favera, and A. Califano, "ARACNE: An Algorithm for the Reconstruction of Gene Regulatory Networks in a Mammalian Cellular Context," *BMC Bioinformatics,* vol. 7, p. S7, 2006.

[74] T. Cover and J. Thomas, *Elements of Information Theory*: New York: John Wiley, 1990.

[75] H. Peng, L. Fulmi, and C. Ding, "Feature selection based on mutual information criteria of max-dependency, max-relevance, and min-redundancy," *IEEE Transactions on Pattern Analysis and Machine Intelligence,* vol. 27, pp. 1226-1238, 2005.

[76] P. E. Meyer, K. Kontos, and G. Bontempi, "Biological network inference using redundancy analysis," in *Bioinformatics Research and Development*: Springer, 2007, pp. 16-27.

[77] J. P. Onnela, K. Kaski, and J. Kertész, "Clustering and information in correlation based financial networks," *The European Physical Journal B - Condensed Matter and Complex Systems,* vol. 38, pp. 353-362, 2004.

[78] S. Bill, *Cause and Correlation in Biology*: Cambridge University Press, 2000.

[79] A. Butte, P. Tamayo, D. Slonim, T. Golub, and I. Kohane, "Discovering functional relationships between RNA expression and chemotherapeutic susceptibility using relevance networks," *Proceedings of the National Academy of Sciences of the United States of America,* vol. 97, pp. 12182 - 12186, 2000.

[80] R. Steuer, "Review: On the analysis and interpretation of correlations in metabolomic data," *Briefings in Bioinformatics,* vol. 7, pp. 151-158, June 1, 2006.

[81] M. Tumminello, T. Aste, T. Di Matteo, and R. N. Mantegna, "A tool for filtering information in complex systems," *Proceedings of the National Academy of Sciences of the United States of America,* vol. 102, pp. 10421-10426, July 26, 2005.

[82] M. B. Eisen, P. T. Spellman, P. O. Brown, and D. Botstein, "Cluster analysis and display of genome-wide expression patterns," *PNAS,* vol. 95, pp. 14863 - 14868, 1998.

[83] J. M. Stuart, E. Segal, D. Koller, and S. K. Kim, "A Gene-Coexpression Network for Global Discovery of Conserved Genetic Modules," *Science,* vol. 302, pp. 249 - 255, 2003.

[84] H. Li, Y. Sun, and M. Zhan, "Exploring pathways from gene co-expression to network dynamics," *Methods in molecular biology (Clifton, N.J.),* vol. 541, pp. 249-267, 2009.

[85] P. Langfelder and S. Horvath, "WGCNA: an R package for weighted correlation network analysis," *BMC Bioinformatics,* vol. 9, p. 559, 2008.

[86] S. Rogers and M. Girolami, "A Bayesian regression approach to the inference of regulatory networks from gene expression data," *Bioinformatics,* vol. 21, pp. 3131 - 3137, 2005.

[87] N. Friedman, "Inferring cellular networks using probabilistic graphical models," *Science,* vol. 303, pp. 799-805, 2004.

[88] A. V. Werhli, M. Grzegorczyk, and D. Husmeier, "Comparative evaluation of reverse engineering gene regulatory networks with relevance networks, graphical Gaussian models and Bayesian networks," *Bioinformatics,* vol. 22, pp. 2523 - 2531, 2006.

[89] X. Chen, M. Chen, and K. Ning, "BNArray: an R package for constructing gene regulatory networks from microarray data by using Bayesian network," *Bioinformatics,* vol. 22, pp. 2952-2954, December 1, 2006.

[90] J. K. Vass, D. J. Higham, M. A. V. Mudaliar, X. Mao, and D. J. Crowther, "Discretization provides a conceptually simple tool to build expression networks," *PLoS ONE,* vol. 6, p. e18634, 2011.

[91] A. Robinson, "Randomization, Bootstrap and Monte Carlo Methods in Biology," *Journal of the Royal Statistical Society: Series A (Statistics in Society),* vol. 170, pp. 856-856, 2007.

[92] T. van den Bulcke, K. van Leemput, and B. Naudts, "SynTReN: a generator of synthetic gene expression data for design and analysis of structure learning algorithms," *BMC Bioinformatics,* vol. 7, p. 43, 2006.

[93] D. Marbach, J. C. Costello, R. Küffner, N. M. Vega, R. J. Prill, D. M. Camacho, K. R. Allison, M. Kellis, J. J. Collins, and G. Stolovitzky, "Wisdom of crowds for robust gene network inference," *Nature Methods,* 2012.

[94] J. D. Allen, Y. Xie, M. Chen, L. Girard, and G. Xiao, "Comparing Statistical Methods for Constructing Large Scale Gene Networks," *PLoS ONE,* vol. 7, p. e29348, 2012.

[95] D. M. Chickering, D. Heckerman, and C. Meek, "Large-sample learning of Bayesian networks is NP-hard," *The Journal of Machine Learning Research,* vol. 5, pp. 1287-1330, 2004.

[96] J. Slawek and T. Arodz, "ENNET: inferring large gene regulatory networks from expression data using gradient boosting," *BMC Systems Biology,* vol. 7, p. 106, 2013.

[97] D. Edwards, *Introduction to Graphical Modelling*: Springer, 2000.

[98] R. Ihaka and R. Gentleman, "R: A Language for Data Analysis and Graphics," *Journal of Computational and Graphical Statistics,* vol. 5, pp. 299-314, 1996.

[99] R. Gentleman, V. Carey, D. Bates, B. Bolstad, M. Dettling, S. Dudoit, B. Ellis, L. Gautier, Y. Ge, J. Gentry, K. Hornik, T. Hothorn, W. Huber, S. Iacus, R. Irizarry, F. Leisch, C. Li, M. Maechler, A. Rossini, G. Sawitzki, C. Smith, G. Smyth, L. Tierney, Y. Yang, and J. Zhang, "Bioconductor: Open software development for computational biology and bioinformatics," *Genome Biology,* vol. 5, 2004.

[100] R. W. Robinson, "Counting labeled acyclic digraphs," *New Directions in the Theory of Graphs,* pp. 239 - 273, 1973.

[101] G. Dennis, B. Sherman, D. Hosack, J. Yang, W. Gao, H. Lane, and R. Lempicki, "DAVID: Database for Annotation, Visualization, and Integrated Discovery," *Genome Biol,* vol. 4, p. P3, 2003.

[102] J. T. Chang and J. R. Nevins, "GATHER: a systems approach to interpreting genomic signatures," *Bioinformatics,* vol. 22, pp. 2926-2933, December 1, 2006.

[103] H. Araki, C. Knapp, P. Tsai, and C. Print, "GeneSetDB: A comprehensive meta-database, statistical and visualisation framework for gene set analysis," *FEBS Open Bio,* vol. 2, pp. 76-82, 2012.

[104] A. Upton and T. N. Arvanitis, "Investigating survival prognosis of glioblastoma using evolutional properties of gene networks," in *2012 IEEE 12th International Conference on Bioinformatics & Bioengineering (BIBE)*, 2012, pp. 466-471.

[105] E. G. Van Meir, C. G. Hadjipanayis, A. D. Norden, H.-K. Shu, P. Y. Wen, and J. J. Olson, "Exciting New Advances in Neuro-Oncology: The Avenue to a Cure for Malignant Glioma," *CA: a cancer journal for clinicians,* vol. 60, pp. 166-193, 2010.

[106] R. G. W. Verhaak, K. A. Hoadley, E. Purdom, V. Wang, Y. Qi, M. D. Wilkerson, C. R. Miller, L. Ding, T. Golub, J. P. Mesirov, G. Alexe, M. Lawrence, M. O'Kelly, P. Tamayo, B. A. Weir, S. Gabriel, W. Winckler, S. Gupta, L. Jakkula, H. S. Feiler, J. G. Hodgson, C. D. James, J. N. Sarkaria, C. Brennan, A. Kahn, P. T. Spellman, R. K. Wilson, T. P. Speed, J. W. Gray, M. Meyerson, G. Getz, C. M. Perou, and D. N. Hayes, "Integrated Genomic Analysis Identifies Clinically Relevant Subtypes of Glioblastoma Characterized by Abnormalities in PDGFRA, IDH1, EGFR, and NF1," *Cancer cell,* vol. 17, pp. 98-110, 2010.

[107] R. A. Irizarry, B. M. Bolstad, F. Collin, L. M. Cope, B. Hobbs, and T. P. Speed, "Summaries of Affymetrix GeneChip probe level data," *Nucleic Acids Research,* vol. 31, p. e15, February 15, 2003.

[108] G. Csardi and T. Nepusz, "The igraph software package for complex network research," *InterJournal Complex Systems,* vol. 1695, 2006.

[109] R. McLendon, A. Friedman, D. Bigner, E. G. Van Meir, D. J. Brat, G. M. Mastrogianakis, J. J. Olson, T. Mikkelsen, N. Lehman, and K. Aldape, "Comprehensive genomic characterization defines human glioblastoma genes and core pathways," *Nature,* vol. 455, pp. 1061-1068, 2008.

[110] G. Gundem, C. Perez-Llamas, A. Jene-Sanz, A. Kedzierska, A. Islam, J. Deu-Pons, S. Furney, and N. Lopez-Bigas, "IntOGen: Integration and data-mining of multidimensional oncogenomic data," *Nature Methods,* 2010.

[111] S. J. Furney, B. Calvo, P. Larrañaga, J. A. Lozano, and N. Lopez-Bigas, "Prioritization of candidate cancer genes - an aid to oncogenomic studies," *Nucleic Acids Research,* vol. 36, p. e115, October 1, 2008.

[112] C. Brennan, H. Momota, D. Hambardzumyan, T. Ozawa, A. Tandon, A. Pedraza, and E. Holland, "Glioblastoma Subclasses Can Be Defined by Activity among Signal Transduction Pathways and Associated Genomic Alterations," *PLoS ONE,* vol. 4, p. e7752, 2009.

[113] H. Dong, L. Luo, S. Hong, H. Siu, Y. Xiao, L. Jin, R. Chen, and M. Xiong, "Integrated analysis of mutations, miRNA and mRNA expression in glioblastoma," *BMC Systems Biology,* vol. 4, p. 163, 2010.

[114] A. M. Donson, A. Banerjee, F. Gamboni-Robertson, J. M. Fleitz, and N. K. Foreman, "Protein Kinase C ζ Isoform is Critical for Proliferation in Human Glioblastoma Cell Lines," *Journal of Neuro-Oncology,* vol. 47, pp. 109-115, 2000.

[115] R. K. Thomas, A. C. Baker, R. M. DeBiasi, W. Winckler, T. LaFramboise, W. M. Lin, M. Wang, W. Feng, T. Zander, L. E. MacConaill, J. C. Lee, R. Nicoletti, C. Hatton, M. Goyette, L. Girard, K. Majmudar, L. Ziaugra, K.-K. Wong, S. Gabriel, R. Beroukhim, M. Peyton, J. Barretina, A. Dutt, C. Emery, H. Greulich, K. Shah, H. Sasaki, A. Gazdar, J. Minna, S. A. Armstrong, I. K. Mellinghoff, F. S. Hodi, G. Dranoff, P. S. Mischel, T. F. Cloughesy, S. F. Nelson, L. M. Liau, K.

Mertz, M. A. Rubin, H. Moch, M. Loda, W. Catalona, J. Fletcher, S. Signoretti, F. Kaye, K. C. Anderson, G. D. Demetri, R. Dummer, S. Wagner, M. Herlyn, W. R. Sellers, M. Meyerson, and L. A. Garraway, "High-throughput oncogene mutation profiling in human cancer," *Nat Genet,* vol. 39, pp. 347-351, 2007.

[116] R. J. Gilbertson, D. A. Hill, R. Hernan, M. Kocak, R. Geyer, J. Olson, A. Gajjar, L. Rush, R. L. Hamilton, S. D. Finkelstein, and I. F. Pollack, "ERBB1 Is Amplified and Overexpressed in High-grade Diffusely Infiltrative Pediatric Brain Stem Glioma," *Clinical Cancer Research,* vol. 9, pp. 3620-3624, September 1, 2003.

[117] C. L. Chang, "Genome-Wide Oligonucleotide Microarray Analysis of Gene-Expression Profiles of Taiwanese Patients with Anaplastic Astrocytoma and Glioblastoma Multiforme," *Journal of Biomolecular Screening,* vol. 13, pp. 912-921, October 1, 2008.

[118] P. J. A. Eichhorn, L. Rodon, A. Gonzalez-Junca, A. Dirac, M. Gili, E. Martinez-Saez, C. Aura, I. Barba, V. Peg, A. Prat, I. Cuartas, J. Jimenez, D. Garcia-Dorado, J. Sahuquillo, R. Bernards, J. Baselga, and J. Seoane, "USP15 stabilizes TGF-β receptor I and promotes oncogenesis through the activation of TGF-β signaling in glioblastoma," *Nat Med,* vol. 18, pp. 429-435, 2012.

[119] S. O. Rahaman, M. A. Vogelbaum, and S. J. Haque, "Aberrant Stat3 Signaling by Interleukin-4 in Malignant Glioma Cells: Involvement of IL-13Rα2," *Cancer Research,* vol. 65, pp. 2956-2963, April 1, 2005.

[120] B. K. A. Rasheed, R. E. McLendon, J. E. Herndon, H. S. Friedman, A. H. Friedman, D. D. Bigner, and S. H. Bigner, "Alterations of the TP53 gene in human gliomas," *Cancer Research,* vol. 54, pp. 1324-1330, 1994.

[121] H. Zheng, H. Ying, H. Yan, A. C. Kimmelman, D. J. Hiller, A.-J. Chen, S. R. Perry, G. Tonon, G. C. Chu, and Z. Ding, "p53 and Pten control neural and glioma stem/progenitor cell renewal and differentiation," *Nature,* vol. 455, pp. 1129-1133, 2008.

[122] A. Y. Chan, S. J. Coniglio, Y. Y. Chuang, D. Michaelson, U. G. Knaus, M. R. Philips, and M. Symons, "Roles of the Rac1 and Rac3 GTPases in human tumor cell invasion," *Oncogene,* vol. 24, pp. 7821-7829, 2005.

[123] J. J. Molenaar, J. Koster, D. A. Zwijnenburg, P. van Sluis, L. J. Valentijn, I. van der Ploeg, M. Hamdi, J. van Nes, B. A. Westerman, and J. van Arkel, "Sequencing of neuroblastoma identifies chromothripsis and defects in neuritogenesis genes," *Nature,* vol. 483, pp. 589-593, 2012.

[124] Q. Wang, S. Diskin, E. Rappaport, E. Attiyeh, Y. Mosse, D. Shue, E. Seiser, J. Jagannathan, S. Shusterman, M. Bansal, D. Khazi, C. Winter, E. Okawa, G. Grant, A. Cnaan, H. Zhao, N.-K. Cheung, W. Gerald, W. London, K. K. Matthay, G. M. Brodeur, and J. M. Maris, "Integrative Genomics Identifies Distinct Molecular Classes of Neuroblastoma and Shows That Multiple Genes Are Targeted by Regional Alterations in DNA Copy Number," *Cancer Research,* vol. 66, pp. 6050-6062, June 15, 2006.

[125] J. M. Maris, "Recent advances in neuroblastoma," *New England Journal of Medicine,* vol. 362, pp. 2202-2211, 2010.

[126] M. L. Schmidt, A. Lal, R. C. Seeger, J. M. Maris, H. Shimada, M. O'Leary, R. B. Gerbing, and K. K. Matthay, "Favorable prognosis for patients 12 to 18 months of age with stage 4 nonamplified MYCN neuroblastoma: a Children's Cancer Group Study," *Journal of Clinical Oncology,* vol. 23, pp. 6474-6480, 2005.

[127] K. Kawa, N. Ohnuma, M. Kaneko, K. Yamamoto, T. Etoh, H. Mugishima, M. Ohhira, J. Yokoyama, F. Bessho, T. Honna, J. Yoshizawa, K. Nakada, M. Iwafuchi, T. Nozaki, J. Mimaya, T. Sawada, T. Nakamura, H. Miyata, K. Yamato, and Y. Tsuchida, "Long-Term Survivors of Advanced Neuroblastoma With MYCN Amplification: A Report of 19 Patients Surviving Disease-Free for More Than 66 Months," *Journal of Clinical Oncology,* vol. 17, pp. 3216-3220, October 1, 1999.

[128] K. De Preter, J. Vandesompele, P. Heimann, N. Yigit, S. Beckman, A. Schramm, A. Eggert, R. Stallings, Y. Benoit, M. Renard, A. Paepe, G. Laureys, S. Pahlman, and F. Speleman, "Human fetal neuroblast and neuroblastoma transcriptome analysis confirms neuroblast origin and highlights neuroblastoma candidate genes," *Genome Biology,* vol. 7, p. R84, 2006.

[129] S. Dedoni, M. C. Olianas, and P. Onali, "Interferon-β induces apoptosis in human SH-SY5Y neuroblastoma cells through activation of JAK-STAT signaling and down-regulation of PI3K/Akt pathway," *Journal of neurochemistry,* vol. 115, pp. 1421-1433, 2010.

[130] A. Iolascon, L. Giordani, A. Borriello, R. Carbone, A. Izzo, G. P. Tonini, C. Gambini, and F. Della Ragione, "Reduced expression of transforming growth factor-beta receptor type III in high stage neuroblastomas," *British journal of cancer,* vol. 82, p. 1171, 2000.

[131] M. C. Tuthill, R. K. Wada, J. M. Arimoto, C. N. Sugino, K. K. Kanemaru, K. K. Takeuchi, and N. Sidell, "N-myc oncogene expression in neuroblastoma is driven by Sp1 and Sp3," *Molecular genetics and metabolism,* vol. 80, pp. 272-280, 2003.

[132] Y. Takei and R. Laskey, "Tumor necrosis factor α regulates responses to nerve growth factor, promoting neural cell survival but suppressing differentiation of neuroblastoma cells," *Molecular biology of the cell,* vol. 19, pp. 855-864, 2008.

[133] M. E. Higgins, M. Claremont, J. E. Major, C. Sander, and A. E. Lash, "CancerGenes: a gene selection resource for cancer genome projects," *Nucleic Acids Research,* vol. 35, pp. D721-D726, January 1, 2007.

[134] S. Álvarez, A. Blanco, M. Fresno, and M. Á. Muñoz-Fernández, "TNF-α Contributes to Caspase-3 Independent Apoptosis in Neuroblastoma Cells: Role of NFAT," *PLoS ONE,* vol. 6, p. e16100, 2011.

[135] C. Kaplinsky, J. Barankiewicz, H. Yeger, and A. Cohen, "PURINE NUCLEOTIDE METABOLISM IN HUMAN NEUROBLASTOMA CELL LINES: 99," *Pediatr Res,* vol. 19, pp. 760-760, 1985.

[136] S. Charrasse, F. Comunale, E. Gilbert, O. Delattre, and C. Gauthier-Rouvière, "Variation in cadherins and catenins expression is linked to both proliferation and transformation of Rhabdomyosarcoma," *Oncogene,* vol. 23, pp. 2420-2430, 2003.

[137] G. Kapatai, M. A. Brundler, H. Jenkinson, P. Kearns, M. Parulekar, A. C. Peet, and C. M. McConville, "Gene expression profiling identifies different sub-types of retinoblastoma," *Br J Cancer,* vol. 109, pp. 512-525, 2013.

[138] A. Ganguly and K. E. Nichols, "Genetics of Retinoblastoma: Molecular and Clinical Aspects," *Retinoblastoma,* p. 24, 2012.

[139] J. Zhang, C. A. Benavente, J. McEvoy, J. Flores-Otero, L. Ding, X. Chen, A. Ulyanov, G. Wu, M. Wilson, J. Wang, R. Brennan, M. Rusch, A. L. Manning, J. Ma, J. Easton, S. Shurtleff, C. Mullighan, S. Pounds, S. Mukatira, P. Gupta, G. Neale, D. Zhao, C. Lu, R. S. Fulton, L. L. Fulton, X. Hong, D. J. Dooling, K. Ochoa, C. Naeve, N. J. Dyson, E. R. Mardis, A. Bahrami, D. Ellison, R. K. Wilson, J. R. Downing, and M. A. Dyer, "A novel retinoblastoma therapy from genomic and epigenetic analyses," *Nature,* vol. 481, pp. 329-334, 2012.

[140] D. Ivanov, G. Dvoriantchikova, L. Nathanson, S. J. McKinnon, and V. I. Shestopalov, "Microarray analysis of gene expression in adult retinal ganglion cells," *FEBS Letters,* vol. 580, pp. 331-335, 2006.

[141] R. W. Young, "Visual cells, daily rhythms, and vision research," *Vision research,* vol. 18, pp. 573-578, 1978.

[142] N. Dyson, "The regulation of E2F by pRB-family proteins," *Genes & development,* vol. 12, pp. 2245-2262, 1998.

[143] M. Reich, T. Liefeld, J. Gould, J. Lerner, P. Tamayo, and J. P. Mesirov, "GenePattern 2.0," *Nat Genet,* vol. 38, pp. 500 - 501, 2006.

[144] A. K. Grennan, "Genevestigator. Facilitating web-based gene-expression analysis," *Plant physiology,* vol. 141, pp. 1164-1166, 2006.

[145] The Bioconductor Development Team, "RGalaxy: Make an R function available in the Galaxy web platform," 1.5.2 ed, 2013.

[146] P. Shannon, A. Markiel, O. Ozier, N. S. Baliga, J. T. Wang, D. Ramage, N. Amin, B. Schwikowski, and T. Ideker, "Cytoscape: a software environment for integrated models of biomolecular interaction networks," *Genome Res*, vol. 13, pp. 2498 - 2504, 2003.

[147] B.-J. Breitkreutz, C. Stark, T. Reguly, L. Boucher, A. Breitkreutz, M. Livstone, R. Oughtred, D. H. Lackner, J. Bähler, V. Wood, K. Dolinski, and M. Tyers, "The BioGRID Interaction Database: 2008 update," *Nucleic Acids Research*, vol. 36, pp. D637-D640, January 1, 2008.

[148] S. Kerrien, Y. Alam-Faruque, B. Aranda, I. Bancarz, A. Bridge, C. Derow, E. Dimmer, M. Feuermann, A. Friedrichsen, R. Huntley, C. Kohler, J. Khadake, C. Leroy, A. Liban, C. Lieftink, L. Montecchi-Palazzi, S. Orchard, J. Risse, K. Robbe, B. Roechert, D. Thorneycroft, Y. Zhang, R. Apweiler, and H. Hermjakob, "IntAct - open source resource for molecular interaction data," *Nucl. Acids Res.*, vol. 35, pp. D561-565, January 12, 2007.

Selected Personal Communication with Experts

[PC1] Email from Dr Carmel McConville, Senior Lecturer in the School of Cancer Sciences at the University of Birmingham, sent on 21st November 2012 at 08:01. This email is regarding the use of the WGCNA network inference method for the inference of multiple categories of gene regulatory network from one microarray dataset. This email specifically details the application to a neuroblastoma dataset, but the issues are the same regardless of the dataset applied to.

Theo,

I've had a look at Alex's data and have put a few comments below which we could discuss.

Best wishes,

Carmel.

I assume that the purpose of the analysis is to find gene networks (with genes within the network showing similar patterns of expression) which might account for the different clinical characteristics of categories 1, 4 and 4M. If a gene shows a change in expression in for example, category 1 compared to category 4, then the whole network might be expected to change in the same direction.

The first thing I did was to look at the lists of genes in the degree distribution, weighted degree distribution and combined distribution for the 3 groups (category 1, category 4, category 4M). Biologically I would expect cat 1 to be very different to the other two, and cat 4 to be somewhat different from cat 4M. Surprisingly there was quite a lot of overlap between categories especially in the combined distribution, where 7 genes in cat 1 were also in either cat 4 or cat 4m.

I next looked at the raw data to try to understand how the analysis was working. From this it was obvious that although the analysis was selecting genes which were highly correlated in their expression, actually expression wasn't differing at all across the 3 categories and this is why the same genes were appearing in different categories. At least some of these genes may have functions which aren't relevant to tumorigenesis.

In addition lots of the selected genes had very low expression (close to background noise) - this is a problem because it's always difficult to know how relevant these are - a two-fold difference in expression might just be 'noise' and likewise apparent correlated expression patterns might be because the genes aren't expressed . We might need to think about doing some additional filtering of the data before the main analysis.

[PC2] Email from Dr Andrew Peet, Reader in Paediatric Oncology in the School of Cancer Sciences at the University of Birmingham and Honorary Consultant at Birmingham Children's Hospital, sent on 21st November 2012 at 08:09. This email is also regarding the use of the WGCNA network inference method for the inference of multiple categories of gene regulatory network from one microarray dataset. This email specifically mentions taking into account the relative expression levels of the genes.

Hi,

I agree that concentrating on the highly expressed genes is a good one, however you probably then loose the closely connected ones which you will need to building networks. How about weighting the genes according to their relative expression levels?

Best wishes,

Andrew