2019

Genet-CNV: Boolean Implication Networks for Modeling Genome-Wide Co-occurrence of DNA Copy Number Variations

Salvi Singh
West Virginia University, ss0083@mix.wvu.edu

Recommended Citation

Singh, Salvi, "Genet-CNV: Boolean Implication Networks for Modeling Genome-Wide Co-occurrence of DNA Copy Number Variations" (2019). *Graduate Theses, Dissertations, and Problem Reports*. 3766.
https://researchrepository.wvu.edu/etd/3766

Genet-CNV: Boolean Implication Networks for Modeling Genome-Wide Co-occurrence of DNA Copy Number Variations

Salvi Singh

Thesis submitted

to the Benjamin M. Statler College of Engineering and Mineral Resources

at West Virginia University

In partial fulfilment of the requirements for the degree of
Master of Science in
Computer Science

Nancy Lan Guo, Ph.D., Chair
Saiph Savage, Ph.D.
Donald Adjeroh, Ph.D.
Lane Department of Computer Science and Electrical Engineering
West Virginia University

Morgantown, West Virginia
2019

Keywords: Boolean Implication Networks, DNA CNV, Genet-CNV

Abstract

Genet-CNV: Boolean Implication Networks for Modelling Genome-Wide Co-occurrence of DNA Copy Number Variations

Salvi Singh

Lung cancer is the leading cause of cancer-related death in the world. Lung cancer can be categorized as non-small cell lung cancer (NSCLC) and small cell lung cancer (SCLC). NSCLC makes up about 80% to 85% of lung cancer cases diagnosed, whereas SCLC is responsible for 10% to 15% of the cases. It remains a challenge for physicians to identify patients who shall benefit from chemotherapy. In such a scenario, identifying genes that can facilitate therapeutic target discoveries and better understanding disease mechanisms and their regulation in different stages of lung cancer, remains an important topic of research.

In this thesis, we develop a computational framework for modelling molecular gene interaction networks, called Genet-CNV, to analyse gene interactions based on DNA Copy Number Variations (CNV). DNA copy number variation is a phenomenon in which sections of the genome are repeated and the number of repeats in the genome varies between individuals in the human population. These variations can be used to study the activity of genes in cancerous cells, compared with that of the normal population. Genet-CNV uses Boolean implication networks to investigate genome-wide DNA CNV to identify relationships called rules, that could potentially lead to the identification of genes of significant biological interest. Boolean implication networks are probabilistic graphical models that express the relationship between two variables terms of six implication rules that can describe if the genes are co-amplified, co-deleted or differentially amplified and deleted. Genet-CNV is run on three publicly available NSCLC genomic datasets. We further evaluate the results obtained with Genet-CNV by comparing them with the benchmark dataset, The Molecular Signatures Database (MSigDB). We identified several genes of interest that are present in survival, apoptosis, proliferation and immunologic pathways. The relationships obtained from this analysis can be tested for biological validations, or to confirm experimental results, thus facilitating the identification of genes playing a significant role in the causation and progress of NSCLC.

Acknowledgements

Dr. Nancy Lan Guo’s exciting work with identifying disease biomarkers to facilitate informed clinical decisions is deeply significant for the understanding and treatment of lung cancer, and thus, for public health in general. In this acknowledgment, I would like to begin by expressing my sincere gratitude to her for providing me with this wonderful opportunity to contribute to her meaningful research, as my supervisor and committee chair. I also wish to thank her for her guidance and assistance at every step, while at the same time pushing me to inculcate sufficient autonomy to be a good researcher. She has benefitted me greatly with her insights and encouragements in all my academic pursuits in general, and this thesis in particular.

I would like to thank Dr. Savage for the unique learning experience she curated for her class, which was an excellent opportunity to learn by implementation. I would also like to thank her for being so generous with her time and ideas whenever I have approached her. She has served as an inspiration for me, both within and outside class. I would also like to express my gratefulness towards Dr. Adjeroh, for his time and guidance as my graduate advisor. It was a privilege to be introduced to Big Data Engineering under his able tutelage.

I would like to thank my family and friends, both back home and here, for their continued love and support. I take great delight in expressing my gratitude towards my roommate, Ritika, and my colleagues at Guo Lab, Rehab and Qing. They have brought much cheer and companionship in my life as a graduate student.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

Chapter 1

Introduction

1.1 Problem and Motivation

Lung cancer is the leading cause of cancer-related death in the world (Torre et al., 2015). Lung cancer can be categorized as non-small cell lung cancer (NSCLC) and small cell lung cancer (SCLC). NSCLC makes up about 80% to 85% of lung cancer cases diagnosed, whereas SCLC is responsible for 10% to 15% of the cases. NSCLC has two major subtypes of histology: squamous cell lung carcinoma and lung adenocarcinoma, with certain DNA mutations causing further molecular stratification (Herbst, Heymach, & Lippman, 2008). NSCLC can have a favorable prognosis if diagnosed at an early stage, with a 5-year survival rate of 70-90% for small localized tumors (stage I) (Goldstraw et al., 2016; Nesbitt, Putnam, Walsh, Roth, & Mountain, 1995; Shah, Sabanathan, Richardson, Mearns, & Goulden, 1996). However, approximately 75% of the patients are diagnosed when they have reached stage III/IV (Walters et al., 2013) and despite significant developments in the oncological management of late stage lung cancer over recent years, survival remains poor (Blandin Knight et al., 2017). SCLC is more aggressive than NSCLC and has a much worse prognosis, with overall 5-year survival around 5% (Blandin Knight et al., 2017).

It remains a challenge for physicians to identify patients who shall benefit from chemotherapy. In such a scenario, identifying genes that can facilitate therapeutic target discoveries and better understanding of disease mechanisms and their regulation in different stages of lung cancer, remains an important topic of research. In this thesis, we develop a computational framework for modelling molecular gene interaction networks, called Genet-CNV, to analyse gene interactions based on DNA Copy Number Variations (CNV) and further investigate genes of significant interest.

DNA copy number variation is a phenomenon in which sections of the genome are repeated and the number of repeats in the genome varies between individuals in the human population. These variations can be used to study the activity of genes in cancerous cells, compared with that of the normal population. The CNV data can be classified into five states, depending upon the number of copies of each gene. The five states are normal, gain, amplification, loss, and double loss.

Implication networks based off of prediction logic were first proposed by Guo et al in 2003 (L. Guo, Cukic, & Singh, 2003). Boolean implication networks have been used to examine gene regulation and control previously (Jansen et al., 2003; Milo et al., 2002; Sachs, Perez, Pe'er, Lauffenburger, & Nolan, 2005; Sahoo, Dill, Gentles, Tibshirani, & Plevritis, 2008). For this study, we are using an implication network based on prediction logic (L. Guo, Cukic, & Singh, 2003). A previous implementation was previously used for genome-wide gene expression modelling by Guo et al in 2011 (N. L. Guo & Wan, 2012). Boolean implication networks are probabilistic graphical models that express the relationship between two variables (genes in this case) in terms of six implication rules that can describe if the genes are co-amplified, co-deleted or differentially amplified and deleted. In prior work, Guo et al have used implication networks to model gene co-expression networks from publicly available datasets. This study extends the previous implication networks implemented in Genet (N. L. Guo & Wan, 2012) to model genome-scale CNV networks. In this study, however, the variable under analysis, instead of limited to being dichotomous, is now allowed to have numerous discrete values, corresponding to CNV states, in logic relations. A software package (Genet-CNV) was developed in this study to model genome-wide CNV networks.

1.2 Thesis Contributions

- Creation of Genet-CNV, a software package developed in C to detect Boolean implication relationships from whole genome DNA CNV data.
- Processing and analysis of three publicly available NSCLC DNA CNV data sets to discover genome wide Boolean implication relationships
- Detection and visualization of results pertaining to seven NSCLC prognostic biomarker genes identified from a previously published study
- Analysis of results to discover results common in datasets, and evaluation of results with the benchmark dataset MSigDB
- Comparison of Genet-CNV's performance on whole genome DNA CNV with algorithms proposed in previous studies

1.3 Thesis Outline

The second chapter presents a brief overview of previous works where Boolean implication networks were employed in genome wide studies. The first section discusses the first application of Boolean implication algorithm to microarray data. The following section describes mining The Cancer Genome Atlas (TCGA) data from various patient cohorts and using Boolean implication networks to generate correlation results. This study looks at gene expression, mutation, methylation and DNA copy number alteration data to discover Boolean implications. It uses a modified version of the algorithm used in the first study that makes up this section. The third section discusses the combined application of self-organizing maps and Boolean implication algorithm to analyse data at various levels of abstraction, including genes, metagenes (representations of similarly expressed genes) and similarly behaving metagene groups (spots).

In chapter 3, we introduce the three datasets used in this study, and describe the processing of the data to prepare it for evaluation with Genet-CNV. We then present the details of the methodology associated with the Boolean implication algorithm. This is followed by a discussion of the development of Genet-CNV to analyse whole genome DNA CNV data for discovering Boolean implication rules in the genome. The performance of genet-CNV is compared with a previously described algorithm.

In chapter 4, we summarize the results obtained from the three datasets. We present the Boolean implication rules that were common across datasets. We further evaluate the results obtained by Genet-CNV against the bench mark dataset called The Molecular Signatures Database (MSigDB), which is a collection of annotated genes that have previously been found to be significant biologically. We preset gene interactions discovered in our datasets and matched with the proliferation, survival, apoptosis and immunology pathways as present in MSigDB. Heatmaps demonstrating clustering of genes present in these four pathways from two of the three datasets have also been presented.

Lastly, in chapter 5, we present our concluding remarks, and the future directions in which this work can be extended for further discoveries.

1.4 Publications Related to this Study

Papers

- Nancy Lan Guo, Afshin Dowlati, Rebecca A. Raese, Chunlin Dong, Guoan Chen, David G. Beer, Justine Shaffer, Salvi Singh, Ujala Bokhary, Lin Liu, John Howington, Thomas Hensing, and Yong Qiane, “A Predictive 7-Gene Assay and Prognostic Protein Biomarkers for Non-small Cell Lung Cancer”, EBioMedicine (N. L. Guo et al., 2018)
- Dymacek JM1, Snyder-Talkington BN, Raese R, Dong C, Singh S, Porter DW, Ducatman B, Wolfarth MG, Andrew ME, Battelli L, Castranova V, Qian Y, Guo NL, “Similar and Differential Canonical Pathways and Biological Processes Associated with Multiwalled Carbon Nanotube and Asbestos-Induced Pulmonary Fibrosis: A 1-Year Postexposure Study.” (Dymacek et al., 2018)
- Brandi N. Snyder-Talkington, Chunlin Dong, Salvi Singh, Rebecca Raese, Yong Qian, Dale Porter, Michael G. Wolfarth, and Nancy L. Guo, “Multi-walled Carbon Nanotube-Induced Gene Expression Biomarkers for Medical and Occupational Surveillance”, Submitted for Publication

Poster

- Salvi Singh, Nancy Lan Guo, “Genet-CNV: Boolean Implication Networks for Modelling Genome-Wide Co-occurrence of DNA Copy Number Variations”, BCB '18 Proceedings of the 2018 ACM International Conference on Bioinformatics, Computational Biology, and Health Informatics

The content of this study shall be divided into two parts, one highlighting the clinical implications of this work, and the other focusing on the algorithm developed and evaluated herein. These are to be submitted for peer-reviewed publication.

Chapter 2

Background and Related Work

2.1 DNA Copy Number Variations (CNVs)

The human genome is made up of 6 billion base pairs of complementary nucleotides, divided into two sets of 23 chromosomes each. The human DNA encodes approximately 30,000 genes. These genes are not always present with two copies in the genome. Large sections of the DNA, ranging from thousands to millions of base-pairs long, can vary in their copy number. They can either have a number of copies which is greater than the norm or have one or more copies lesser than the norm. Sometimes these sections of the DNA can correspond to genes, in which case, it might interfere with the normal functioning of the genes in the genome. Such sections of the DNA comprise Copy Number Variations (CNVs) and are defined as structural repeats in the genome that are larger in size than 1 kilobase (kb).

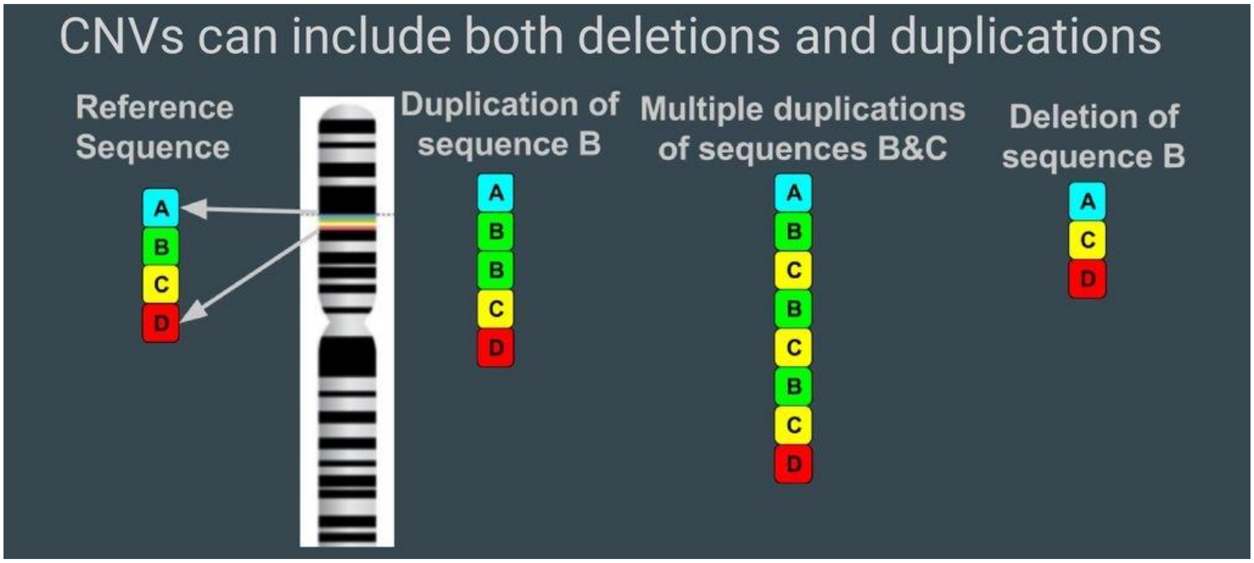

CNVs can include both deletions and duplications
Reference Sequence
A
B
C
D
Duplication of sequence B
A
B
B
C
D
Multiple duplications of sequences B&C
A
B
C
B
C
B
C
D
Deletion of sequence B
A
C
D

Figure 3.1 DNA Copy Number Variation: Deletion and Amplification

While CNVs exist in healthy members of the population, they can still be linked to certain genetic mutations and diseases. Most CNVs are relatively benign, however, sometimes they may affect important developmental genes in the genome and thus cause diseases. Many studies have been conducted which have identified certain CNVs having an association with multiple diseases. It has also been observed that CNVs can encompass multiple genes, and that the effect of a potentially disease-causing CNV is not restricted to the genes it includes and may affect other genes as well. Genome wide studies of CNV datasets are carried out to identify more such CNVs and their potential implications on the rest of the human genome and human health in general. The genomic instability and structural dynamism of cancer cells necessitates that CNV data be examined in order to discover any underlying associations. The prevalence of next-generation DNA microarray-based technologies generate ample data and facilitate the detection of CNVs and further analysis of correlations in the data.

In this study, we have used genome wide non-small cell lung cancer (NSCLC) DNA CNV data and firstly converted it to copy number calls using specialised packages that help generate copy number variation values for each probe in the dataset. Once the calls have been generated, then a Boolean implication network has been created to detect and analyse correlation amongst genes. Copy number variations have thus been used to generate gene correlations which may play a role in identifying genes significant in the cause and course of NSCLC.

2.2 Network-based Methods

Molecular network analysis using computational network models has led to promising applications in identifying new disease genes (Emilsson et al., 2008), discovering disease-related sub-networks (Calvano et al., 2005), and classifying diseases (Chuang, Lee, Liu, Lee, & Ideker, 2007). In this study, we are implementing a novel computational network model using microarray data to analyse gene association relationships, using Boolean Implication Networks. In this section, we discuss and compare two other network models that can be used to carry out a similar analysis.

2.2.1 Artificial Neural Networks [ANN]

ANNs are inspired by the biological nervous system and comprise of several highly interconnected nodes. These nodes are organized into a minimum of three layers, the input layer, the output layer, and the hidden layer. Often, ANNs will have more than one hidden layer. The hidden layers are where the processing is done. The nodes and edges have weights and biases attached to them, which are adjusted when the algorithm is run, to obtain a function that best describes the data.

The backpropagation algorithm is the most commonly used method to iterate over the data and adjust the weights and biases till the error rate between the observed and the expected value has been minimized (Mitchell, 1997).

Neural networks, however, are very complex often acting like black-boxes, and it is very difficult to determine the significance of weights and biases in a biological context, or to compare it with other rule-based methods (Mitchell, 1997). Neural networks also need a huge amount of trained data, in order to successfully model the underlying function. With smaller sample sizes, they lead to over-fitting of the data, and the results do not generalize well to new inputs.

2.2.2 Bayesian Belief Networks

Bayesian Networks are Directed Acyclic Graphs (DAGs), where each node represents a hypothesis or a random variable, and the edges represent direct relationships between the variables. They model the joint probability distributions of random variables. The conditional dependence relationships of the variables are based upon Bayesian probabilities.

Bayesian networks have been used commonly in molecular network analysis, due to their ability to provide causal relationships between pairs of genes, specifically to predict genome wide protein-protein interactions and model cellular networks (Jansen et al., 2003). The Bayesian structure, however, cannot always yield causal relationships. Bayesian networks, being acyclic in nature, cannot accommodate feedback loops (Sachs et al., 2005).

2.4 Studies Using Boolean Implication Networks Algorithm with Cancer Data

2.4.1 Boolean implication networks derived from large scale, whole genome microarray datasets (Sahoo et al., 2008)

This study was the first attempt to discover Boolean implications for the full genome on large-scale gene expression data. Boolean implication rules can be used to identify a larger set of relationships between pairs of genes across the whole genome using data generated from microarray experiments. In this study done using microarray data, the gene expression level of each gene on each of the arrays has been classified as low or high, depending on a threshold determination method applied to each gene. The identified implication rules make up a labeled directed graph, where the vertices are the genes, and the edges are the implication rules, labeled with their type. The authors mention that a Boolean implication relation is an empirical observation and does not imply any causality. However, it can serve as a starting point to examine certain gene interactions and their biological significance.

The study also found that Boolean implications capture a number of relationships that are not identified by other existing methods for large scale data, and most of these methods find only symmetric relationships, whereas with Boolean implication, asymmetric relationships can also be identified. There may exist significant Boolean relationships between genes whose expression is not very strongly correlated. In their study, the authors discovered that the network identifies Boolean implications that describe known biological interactions, and many new relationships that can be used to generate new hypotheses. Many of the new relationships identified in humans by this methodology are common across humans, mice, and fruit flies.

A comparative study on a small set of biologically interesting genes in humans was done to examine the properties of Boolean implications networks with correlation-based networks. Boolean networks captured more rules. The symmetric Boolean relationships almost always had

Publicly available microarray data from humans, mice and fruit-flies was used in this study, with approximately 3 billion probe set pairs examined for possible Boolean implication rules in the human dataset itself. A large number of Boolean implications were found for each individual species.

Some of the findings of this study included the appearance of implication between genes expressed during differentiation of specific tissue types. In the human network, hundreds of genes related to the cell-cycle were found to be co-expressed. Several Boolean implications were also discovered amongst developmentally regulated genes. Many Boolean implication relationships were conserved across multiple species, with about 41,000 of the identified relationships being common across all three species.

The StepMiner (Sahoo, Dill, Tibshirani, & Plevritis, 2007) algorithm was used to determine a threshold level for each gene, and depending on the threshold, the genes were classified as having high expression levels or having low expression levels. Scatter plots were generated for each gene pair for visualization. Depending on the sparsity of the quadrants, Boolean implications were discovered. The number of expression values in a sparse quadrant must be less than the number expected under an independence model. This was followed by the calculation of the error rate in each case. If the first statistic was greater than 3.0 and the error rate less than 0.1, then the implication would be deemed significant.

2.4.2 Mining TCGA Data Using Boolean Implications (Sinha 2014)

In this study, Boolean implications are applied to find relationships between variables of different data types, including mutation, copy number alteration, DNA methylation, and gene expression. The data in this study has been taken from glioblastoma (GBM) and ovarian serous cystadenoma (OV) data sets from The Cancer Genome Atlas (TCGA). Several hundred thousand Boolean implications were discovered in these datasets.

The GBM dataset had 126 patients with mutation and copy number data, 235 with methylation and expression data and 86 patients with mutation, copy number and expression data. The OV dataset used has 314 patients with mutation and copy number data and 286 patients with mutation, copy number, methylations and expression data.

The different data types (gene expression, methylation, copy number alteration, mutation) needed to be discretized into Boolean values in order to discover Boolean implications amongst them. This was the preliminary step before pairwise association relationships could be extracted. The process of extracting Boolean implications between gene pairs involved two major steps: firstly, the chi-square test for independence was used to discover nonrandom associations, and secondly, for the cases where nonrandom associations were found, the sparsity test mentioned in the previous related work was used to produce the final results. An implication was considered significant if the first statistic was greater than a cutoff threshold (typically, between 2.0 and 3.0) and the error rate was less than 0.1.

The chi-square independence test does not generate a good measure for association for when any of the categorical variables have a frequency less than 5. Therefore, in this study, the Fisher's Exact test was used for detection of nonrandom associations, followed by the sparsity check test, for low frequency events such as mutations and copy number alterations.

On further examining the Boolean implications discovered by this process, it was found that these rules accurately capture several known biological phenomena, such as *cis*-regulatory mechanisms of gene regulation, temporal ordering, interactions of multiple pathways, loss-of-heterozygosity of tumor suppressors and mutation-specific epigenetic states. Many of these relationships were found between different data types of the same gene pairs. Further analysis using GSEA showed that the genes obtained by these Boolean implications were biologically meaningful and had overlaps with known cancer genes. The authors mention that the plethora of diverse biological insights generated by examining these relationships indicates that Boolean implications are a very useful tool for mining relationships from cancer data sets to gain further insights.

The authors also compared the Boolean implications relationships generated from the data with relationships discovered by three other techniques used commonly to find pairwise association. These three techniques included the t test, correlation, ad Fisher's exact test. It was found that Boolean implications resulted in the maximum number of relationships out of all the four methods applied. The most significant n rules generated by each of the three other statistical methods were compared with the most significant n rules generated by Boolean implications, and it was found in all three cases that there was very little overlap between Boolean relationships and the relationships generated by any one of the three other methods. The authors therefore concluded that the relationships found by Boolean implications were unique and would be missed by other methods.

In conclusion, the authors state that Boolean implications can be used to explore large datasets and expose numerous symmetric and asymmetric Boolean relationships, which can then lead to new hypotheses and novel biological insights. A potential future application identified in this study is to generate high-order relationships for biologically meaningful variable combinations. In general, apart from generating new result for formulating new hypotheses, relationships found by Boolean implications can be used in conjunction with other data to investigate specific biological questions.

2.4.3 Profiling of Genetic Switches using Boolean Implications in Expression Data (Çakır, 2016)

In this study, a combined approach is used where SOM (self-organizing maps), a machine learning methodology and Boolean implication network are used to identify relations between genes at various levels of abstraction, including genes, metagenes (representations of similarly expressed genes) and similarly behaving metagene groups (spots). Boolean implications can discover weakly correlated entities as well and categorize the relationships obtained into six different types. According to the authors, this method allows them to validate and identify various potential relationships between genes and functional modules of interest, and their switching behavior. This methodology allows the construction and analysis of the network of genes.

The data used in this study was obtained from a publicly available microarray study on 221 mature aggressive B-cell lymphomas. The preliminary step once again entails obtaining gene expression data and converting it into discretized high and low values. For this purpose, the StepMiner (Sahoo et al., 2007) algorithm devised by Sahoo et al (Sahoo et al., 2008) in the first work is utilized. The Boolean implication relationships are calculated using the same workflow and thresholds as used in the first related work by Sahoo et al (Sahoo et al., 2008).

The SOM algorithm is used to transform the expression matrix comprising of thousands of probes representing genes and their expression values for each patient sample into meta-data of reduced dimensionality. This means that groups of genes having similar expression profiles are organized into meta-genes, which are independent cluster representations of these similar genes. No primary information is lost, as none of the features are precluded from the analysis. After this, the expression state of each sample is visualized by color-coding two-dimensional grids of metagenes according to their expression values in the respective samples. This generates colour-gradient map for each sample; since the order of the metagenes is the same for each sample, the expression profiles in case of each sample can be compared directly with that of another sample. Individual expression patterns emerge as spots of similar colored tiles in these color-gradient maps. These spots are groups of co-expressed metagenes for each sample. The two percent metagenes that show the highest and the lowest expression levels respectively for each sample are kept for analysis.

An important consideration in this study was at what level is the extraction of Boolean implications relationships most beneficial. There were three possible pathways of executions identified and compared with each other. In the first approach, the SOM method can be utilized to aggregate single genes to metagenes and further to detect spot clusters and the corresponding spot expression profiles. After that, spot level Boolean implications could be detected. Secondly, the metagenes obtained from SOM are used for pairwise implication analysis, and then spot relationships are discovered from these. And the third approach could be where Boolean implications are found first for the single genes and there pairwise relationships are mapped and aggregated to obtain metagenes and spots.

After comparing these three approaches, this study found that obtaining Boolean implications at the lowest level of single gene-pairs conserved the most information by giving the greatest number of relationships. Under the other two approaches, the number of significant relationships was much fewer. Therefore, the detection of implication relationships, followed by their aggregation to obtain metagene and spot relationships should be used.

This study concludes that the metagenes and spot profiles generated by SOM provide an overview of the data at different levels and reduce the dimension of the data, thus avoiding being bottlenecked by invariant genes present in the data. The uninteresting elements are thus filtered out, and the Boolean implication relationships detected are of a more significant nature. The resulting implication relationships yield a network of genes where the genes/metagenes/spots make up the nodes, while the edges indicate the implication relationship amongst them, which is not possible in case of other methods, such as correlation-based networks. Also, unlike correlations-based networks, not only symmetric but asymmetric relationships are also captured via Boolean networks.

The output of Boolean implication analysis are logical relations of pairs that in turn provides a network of implications where genes/metagenes/spots constitute the nodes and edges stand for the relations between genes/metagenes/spots. Essentially resultant implication networks have a different structure than correlation networks because of their six different edge types. Besides, given a temporal ordering in the feature set of the data an implication network is a directed network i.e. an implication is not necessarily a two-way logical relation.

In the future, Boolean implication networks can be used as modelling approaches for gene states and their interactions. Furthermore, an analysis of the Boolean implications of the genes that take part in lymphoma in a given data set can lead to estimation of the biological fate of the system and identify processes that that can be manipulated and adjusted to observe desired results. This approach can prove useful in the fields of cancer biology, systems biology and clinical applications.

Chapter 3

Genet-CNV: Boolean Implication Networks for Modeling Genome Wide Co-Occurrence of DNA CNV

3.1 Boolean Implications Networks

A probabilistic graphical model is a graph structure where each node of the graph is a representation of a random variable and the edges encode the relationships between these variables. Based on whether the graph is directed or undirected, there are two major types of probabilistic graphical models: Bayesian model (directed) and Markov model (undirected).

Implication networks are also probabilistic graphical models, where each node represents a variable, and the connecting edge between any two nodes describes the implication relationship between them. One major advantage that implication networks have over Bayesian models is that they can be used to represent cyclic relationships as well, whereas Bayesian networks are acyclic in nature.

The first formalism of implication networks was proposed by Liu and Desmarais (Liu and Desmarais, 1997) and was based on binomial distribution.

3.2 Implementation by Sahoo et al (Sahoo et al., 2008)

Another form of implication networks, Boolean implication networks, were constructed by Sahoo et al to model gene interactions networks in a meta-analysis of microarray data for multiple species. The implication relations in the Boolean implication networks were induced based on scatter plots of expression between two genes. On the scatter plots of gene expressions, a threshold was automatically determined using StepMiner (Sahoo et al., 2007) algorithm to discretize the gene expression level as 'high' or 'low'. Based on the discretized levels, the scatter plot is partitioned into four quadrants and the implication relation between the two genes is derived based on the number of data points in the quadrants. In the partitioned scatter plot with four quadrants, the 'low' and 'high' expression of gene A corresponds to $\neg A$ and A respectively. In order to derive a successful implication rule between the pair of genes for the Boolean implication networks, two statistics were tested. The first statistic tests if the observed number of occurrences in the sparse quadrant (error cell) is significantly less than the expected number of occurrences under an independent model, given the relative distribution of low and high values of both genes. The second statistic estimates the maximum likelihood of the error rate for the number of occurrences in the error cell. More details about this study have been described in the related works section.

3.3 Boolean Implication Networks Based on Prediction Logic

Implication networks based off of prediction logic were proposed by Guo et al. In this methodology, implication rules were obtained by using prediction logic based on formal logic rules. There can be six possible implication rules between any pair of dichotomous variables, as described in the figure that follows. In the figure, all shaded cells represent the error cells for each rule.

The first rule is that of positive implication, where if A is true, then B is also true. The error in this case would be if A were to be true but B false. The second rule is of forward negative implication, where if A is true, then B is not true; the error case being if A is true, but B is also true. The third rule is that of inverse negative implication, where if A is not true, then B is true; here the error occurs if when A is not true, B is also not true. The fourth rule is negative implication, if A is not true, then B is also not true; the error occurring when if A is not true, but B is true.

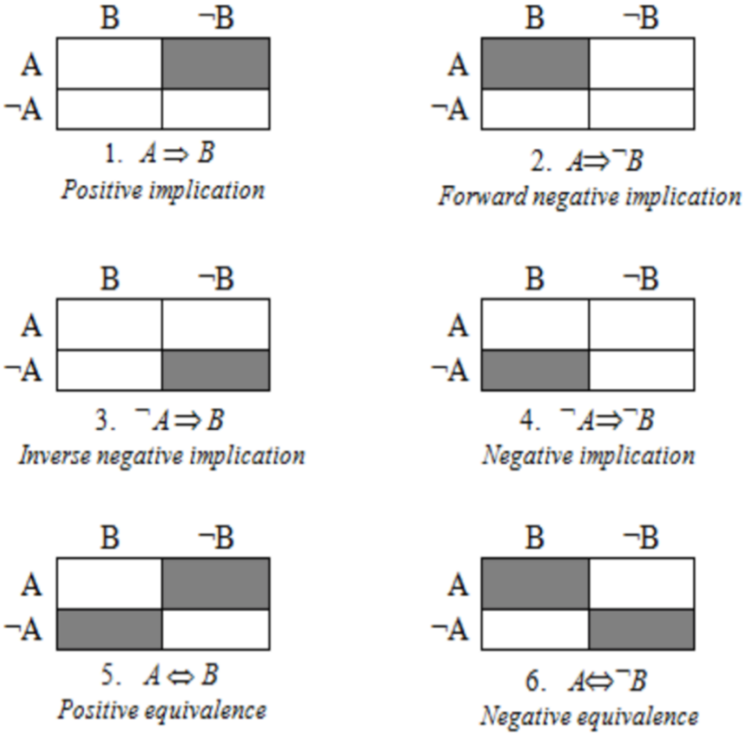

B
¬B
A
¬A
1. $A \Rightarrow B$
Positive implication
2. $A \Rightarrow \neg B$
Forward negative implication
3. $\neg A \Rightarrow B$
Inverse negative implication
4. $\neg A \Rightarrow \neg B$
Negative implication
5. $A \Leftrightarrow B$
Positive equivalence
6. $A \Leftrightarrow \neg B$
Negative equivalence

Figure 3.2 Six implication rules relating two dichotomous variables (shaded cells represent error)

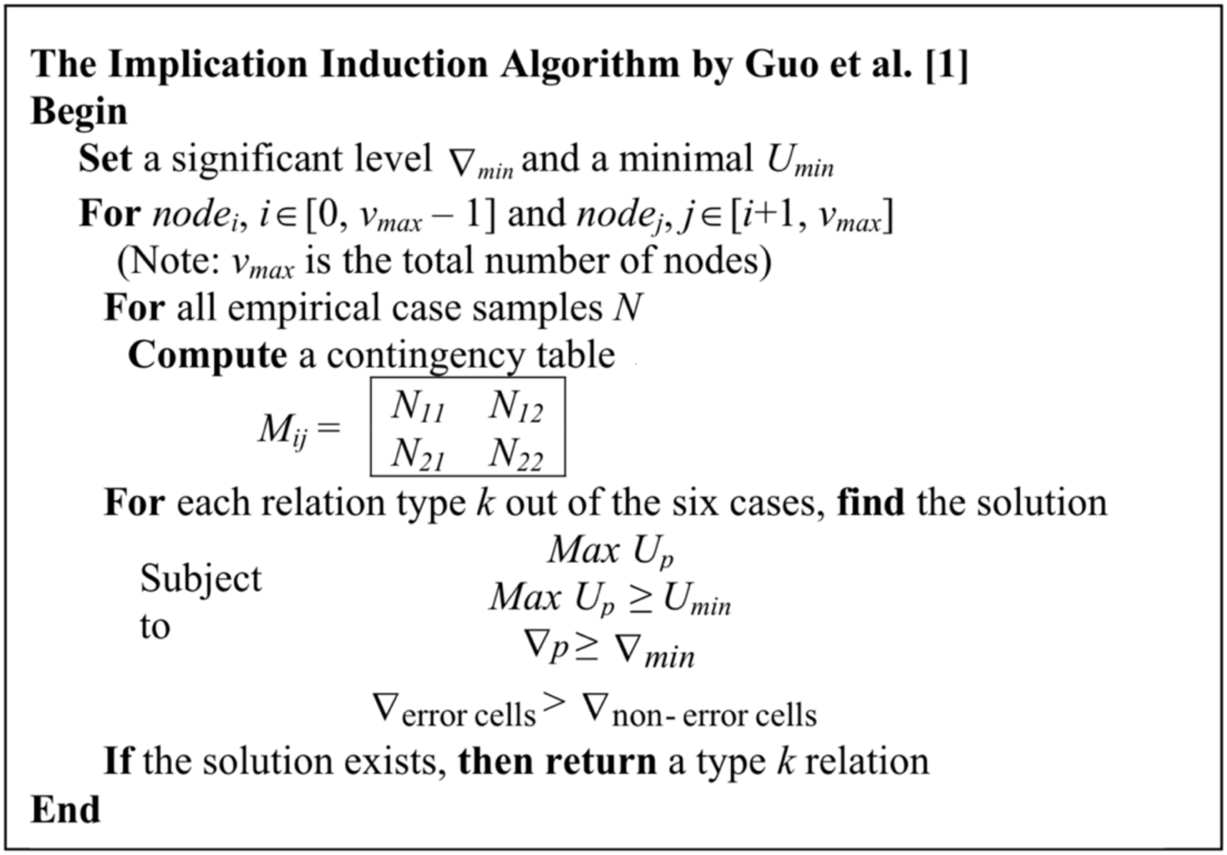
	B	¬B
A	$N_{A\wedge B}$	$N_{A\wedge\neg B}$
¬A	$N_{\neg A\wedge B}$	$N_{\neg A\wedge\neg B}$

Figure 3.3 Contingency table of two variables for *N* empirical samples.

The fifth and sixth rule presented in the Fig 3.2 are equivalence rules. Rule 5, or positive equivalence occurs when rule 1 and rule 4 are true, i.e. if A is true, then B is true, and if B is true, then A is true. The error in this case comprises cases where if A is true, but B is not, and also if B is true, but A is not. The sixth rule is that of negative equivalence and is applicable when both rule 2 and rule 3 are valid, i.e. if A is true, then B is not true, and when A is not true, B is true. The error in this case is counted when A is true and B is also true, and also if A is not true and B is not true.

In the contingency table in Fig 3.3, each cell represents the number of co-occurrences for a particular implication. For example, cell $N_{A\wedge B}$ indicates the number of samples where both variables A and variable B are true. The shaded cells of the contingency table represent the **errors** for the corresponding implication rule. For example, **A∧¬B** is the **error cell** for the **implication rule A ⇒ B**, **$N_{A\wedge\neg B}$** represents the **number of error occurrences**. Cell A∧¬B is erroneous for the rule A ⇒ B because in an ideal case, if the implication A ⇒ B is the true relationship between A and B, then we would never expect to find the contradiction case where A is true but not B. It is to be noted that in the context of DNA copy number data, the states of true and false default to amplified and deleted.

In the original implication induction algorithm, a modified U-optimality method was used to obtain the implication rules between each pair of variables:



```
The Implication Induction Algorithm by Guo et al. [1]
Begin
  Set a significant level $\nabla_{min}$ and a minimal $U_{min}$
  For $node_i$, $i \in [0, v_{max} - 1]$ and $node_j$, $j \in [i+1, v_{max}]$
   (Note: $v_{max}$ is the total number of nodes)
    For all empirical case samples $N$
      Compute a contingency table
        $M_{ij} = \boxed{\begin{matrix} N_{11} & N_{12} \\ N_{21} & N_{22} \end{matrix}}$
    For each relation type $k$ out of the six cases, find the solution
        $Max\ U_p$
      Subject to
        $Max\ U_p \geq U_{min}$
        $\nabla p \geq \nabla_{min}$
        $\nabla_{\text{error cells}} > \nabla_{\text{non-error cells}}$
    If the solution exists, then return a type $k$ relation
End
```

Figure 3.4 Implication induction algorithm based on prediction logic

In the contingency table M_{ij} of the induction algorithm, N_{11} indicates number of samples where both *i* and *j* occur to be true, N_{12} is when *i* is true but not *j*, N_{21} is when *j* is true but not *i*, and N_{22} is when both *i* and *j* are not true.

In the induction algorithm, U_p is the **scope** of the implication rule, representing the **portion of the data covered by the implication relation**, and ∇_p is the **precision** of the implication rule, **representing the prediction success of the corresponding implication relation**. For a single error cell, where N_{ij} is the number of error occurrences, the scope U_p and the precision ∇_p are defined as:

Scope:

$$U_p = U_{ij} = \frac{N_{i\cdot} * N_{\cdot j}}{N^2}$$

Precision:

$$\nabla_p = \nabla_{ij} = 1 - \frac{N_{ij}}{N * U_p}$$

For the rule types where there are multiple error cells, they are defined as:

Scope:

$$U_p = \sum_{\mathrm{i}} \sum_{j} \omega_{\mathrm{ij}} * U_{ij}$$

Precision:

$$\nabla_{\mathrm{p}} = \sum_{\mathrm{i}} \sum_{j} \left(\frac{\omega_{\mathrm{ij}} * U_{ij}}{U_p} \right) \nabla_{\mathrm{ij}}$$

where ω_{ij} = 1 for error cells; otherwise, ω_{ij} = 0.

Based on the contingency table for variable *A* and *B* (M_{AB}) (Figure 3.3), the scope and precision for each of the six implication rules in Fig. 3.2 are defined as follows.

For positive implication, $A \Rightarrow B$,

$$U_{A \Rightarrow B} = U_{A \wedge \neg B} = \frac{N_A * N_{\neg B}}{N^2}$$
$$\nabla_{A \Rightarrow B} = \nabla_{A \wedge \neg B} = 1 - \frac{N_{A \wedge \neg B}}{N * U_{A \Rightarrow B}}$$

Similarly, for forward negative implication, $A \Rightarrow \neg B$,

$$U_{A \Rightarrow \neg B} = U_{A \wedge B} = \frac{N_A * N_B}{N^2}$$
$$\nabla_{A \Rightarrow \neg B} = \nabla_{A \wedge B} = 1 - \frac{N_{A \wedge B}}{N * U_{A \Rightarrow \neg B}}$$

For inverse negative implication, $\neg A \Rightarrow B$,

$$U_{\neg A\Rightarrow B} = U_{\neg A\wedge\neg B} = \frac{N_{\neg A} * N_{\neg B}}{N^2}$$
$$\nabla_{\neg A\Rightarrow B} = \nabla_{\neg A\wedge\neg B} = 1 - \frac{N_{\neg A\wedge\neg B}}{N * U_{\neg A\Rightarrow B}}$$

For negative implication, ¬ *A*⇒ ¬*B*,

$$U_{\neg A\Rightarrow\neg B} = U_{\neg A\wedge B} = \frac{N_{\neg A} * N_{B}}{N^2}$$
$$\nabla_{\neg A\Rightarrow\neg B} = \nabla_{\neg A\wedge B} = 1 - \frac{N_{\neg A\wedge B}}{N * U_{\neg A\Rightarrow\neg B}}$$

For positive equivalence, *A* ⇔ *B,*

$$U_{A\Leftrightarrow B} = U_{A\wedge\neg B} + U_{A\wedge\neg B} = \frac{N_A * N_{\neg B} + N_{\neg A} * N_B}{N^2}$$
$$\nabla_{A\Leftrightarrow B} = 1 - \frac{N_{A\wedge\neg B} + N_{\neg A\wedge B}}{N_A * N_{\neg B} + N_{\neg A} * N_B} * N$$

And for negative equivalence, *A* ⇔ ¬*B*,

$$U_{A\Leftrightarrow\neg B} = U_{A\wedge B} + U_{\neg A\wedge\neg B} = \frac{N_A * N_B + N_{\neg A} * N_{\neg B}}{N^2}$$
$$\nabla_{A\Leftrightarrow B} = 1 - \frac{N_{A\wedge B} + N_{\neg A\wedge\neg B}}{N_A * N_B + N_{\neg A} * N_{\neg B}} * N$$

In the implication induction algorithm, the minimum requirements for the scope (U_{min}) and precision (∇_{min}) must be positive values for an implication rule. They are the parameters used to control the significance level for an implication rule. In our study, we defined the minimum requirement for these two parameters to be at least 95% significant ($P < 0.05$) from one-tail Z- test based on the sample size. In this induction algorithm, a threshold is established individually for both precision and scope, whereas in the original U-Optimality method, a minimum requirement was established for precision alone.

3.4 Genet-CNV

In this study's implementation, called Genet-CNV, we are using Boolean implication networks to examine implication relationships between genes that look at co-amplification, co-deletion, amplification-deletion, and deletion-amplification. The Boolean implication algorithm uses the error cells' count for a particular rule while calculating the scope and precision for that rule. This is a key factor in our approach which enables us to observe relationships between two genes despite the fact that for a given gene, each sample can assume not two but multiple values, representing the CNV state for that sample. For instance, while looking for co-amplification in a pair of genes ($A \Rightarrow B$), Genet-CNV checks if a gene is amplified, and the remaining two states of deletion and normalcy become simply the negate of amplification; the error would be when gene A is amplified but gene B is not amplified. In this manner, Genet-CNV can use the same fundamental principle to create implication relations amongst genes, irrespective of the number of states for each gene.

Genet-CNV allows the user to tune the network in several ways to reach predefined scope and precision threshold, and statistical significance of each implication rule. The threshold values of scope and precision are calculated separately using one-tailed Z-test based on the sample size. All the results presented in this study were obtained at a 95% significance level ($P < 0.05$), unless specified otherwise. The threshold for scope and precision and significance level can be tuned as needed by the user.

The mean precision and scope values have been obtained, they can be used in the test for significant threshold. Alternatively, the user can also use an arbitrary mean for precision and scope, depending on how stringently significant they want the Boolean implications obtained to be. The higher the mean value of precision and scope used, the lower is the number of the resulting significant rules.

After the mean has been decided upon, and the significance level set, the actual Boolean implication detection portion of the algorithm runs on the dataset and produces the number of rules and their types obtained amongst the interacting genes.

We ran Genet-CNV four times on each dataset, each time focusing on one of the four possible interaction types: amplification-amplification, deletion-deletion, amplification-deletion and deletion-amplification.

The maximum time taken by Genet-CNV to analyse a dataset was 110 minutes, on a system running the 64-bit version of Windows 10 Enterprise OS, with an Intel® Xeon 3104 CPU @ 1.70 GHz, and 16.0 GB RAM. This was in the case of whole genome comparison with whole genome for dataset GSE31800, yielding approximately 33 million rules. When discovering implication rules for seven genes with the rest of the genome, Genet-CNV would typically take less than a minute on the same configuration.

3.5 Workflow for Generating Rules and Analysis

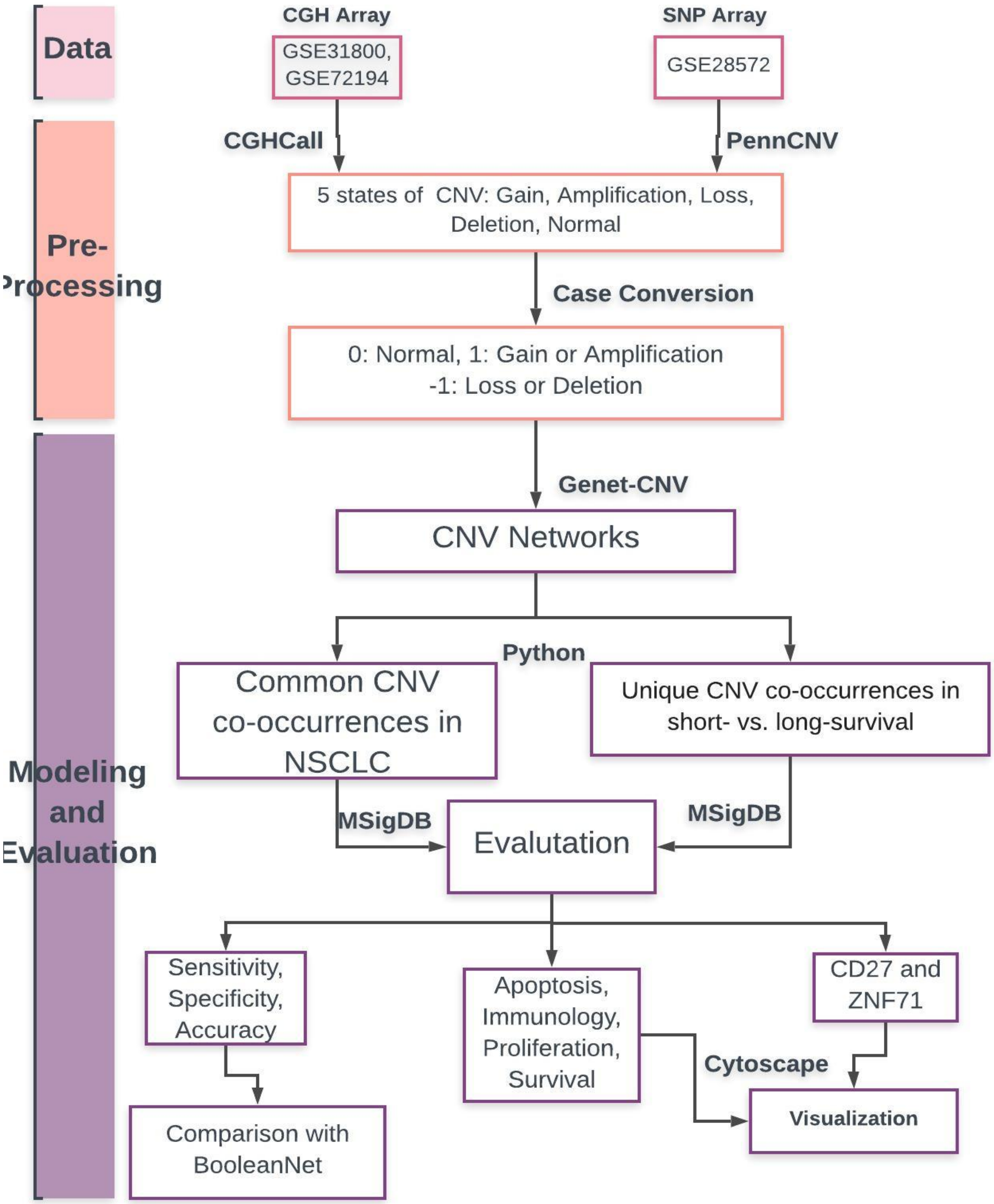


Figure 3.5 Workflow: Data, Pre-processing, Modeling and Evaluation

3.6 Datasets

There were two different types of data used in this study, CGH Microarray data and SNP Microarray data. The first two datasets, GSE31800 and GSE72194, were generated using CGH microarrays, whereas the third dataset created on a SNP microarray platform.

CGH Microarray Data [GSE31800, GSE72194]

Three datasets were used in this study. The first dataset contains 271 NSCLC tumor samples with DNA copy number profiles. These 271 samples are histologically divided into lung AC (*n* = 179) and squamous cell carcinoma (SQCC; *n* = 92). This dataset also contains GE profiles for 49 samples (*n* = 29 for AC; *n* = 20 for SQCC). The patient data was collected using Custom Rosetta-Affymetrix Human platform. Two-channel microarrays were used, with lung tumor tissue analyzed in channel one and normal tissue used as reference in channel two. To obtain the final intensity values, the data was normalized using the algorithm described by Khojasteh et al (Khojasteh, Lam, Ward, & MacAulay, 2005). This dataset is available in NCBI Gene Expression Omnibus (GEO) with accession number GSE31800 (Starczynowski et al., 2011).

The second dataset contains 64 NSCLC tumor samples with DNA copy number profiles. The samples are collected from 64 early stage Non-Small Cell Lung Cancer (NSCLC) patients, which includes lung AC (n = 50), and lung SQCC (n = 14). The patient data was analyzed using Agilent-014693 Human Genome CGH Microarray 244A. Two-channel microarrays have been used for this analysis. The data for every sample was normalized and the intensity was background corrected, in order to obtain log2 ratios for each probe. This dataset is also publicly available in NCBI Gene Expression Omnibus (GEO) with accession number GSE72194 (Aramburu et al., 2015).

DNA copy number profiles of the first dataset were quantified for each sample with whole-genome tiling path array comparative genomic hybridization (aCGH). aCGH is a technique for measuring the changes in chromosomal segments (Solinas-Toldo et al., 1997). The test and reference DNAs are differentially fluorescent labeled and hybridized together to the array in aCGH. The resulting fluorescent ratio is then measured, clone by clone, and plotted relative to each clone's position in the genome (Carter, 2007).

DNA copy number profiles for the second dataset were quantified using Agilent's Oligonucleotide Array-Based CGH for Genomic DNA Analysis protocol, which is also an aCGH technique.

Both GSE31800 and GSE72194 use two channel microarrays, which are hybridized with cDNA from two samples to be compared, for example, diseased tissue and healthy tissue. These two types of tissues are colored using two different fluorescent dyes, Cy3 and Cy5, where Cy3 corresponds to the green part of the spectrum and Cy5 corresponds to the red part.

SNP Microarray Data [GSE28572]

The third dataset comprises a 101 NSCLC patient samples, divided into short-term survivors (<20 months; n=53) and long-term survivors (>58 months; n=47), and one normal sample. The 100 samples are histologically divided into lung adenocarcinoma (n = 51), squamous cell carcinoma (n = 28) and large cell carcinoma (n = 21). The dataset was analyzed on Affymetrix Mapping 250K Nsp SNP Array, and protocols corresponding to this platform were used for pre-processing. The log2 ratios were obtained by using the Copy Number Analysis Tool (CNAT). This dataset is available in NCBI Gene Expression Omnibus (GEO) with accession number GSE28572 (Micke et al., 2011).

Data for the third sample set was collected by using Genome variation profiling by single nucleotide polymorphisms (SNP) arrays. SNPs represents a difference in a single DNA building block, i.e. a nucleotide, resulting in a sequence variation at the single nucleotide. For example, a SNP may replace the nucleotide cytosine (C) with the nucleotide thymine (T) in a certain stretch of DNA. SNPs occur normally throughout a person’s DNA. They occur almost once in every 1,000 nucleotides on average, which means there are roughly 4 to 5 million SNPs in a person's genome. SNPs can serve as markers for various genomic phenomena and high-throughput array technologies are used for SNP genotyping (Genetics Home Reference, 2019, February 3).

The table on the following pages summarize the clinical information available for the patient cohorts analyzed in this study. The table describes the clinical information for GSE28572, where the patient samples are further categorized according to their survival status. It also presents a summary of the clinical information for the two CGH microarray datasets, GSE72194 and GSE31800 respectively.

Table 3.1 Clinical Information of Patient Samples from GSE31800, GSE72194 and GSE28572

VARIABLES	**GSE31800, n = 271**	**GSE72194, *n = 64***	**Long-Term Survivors (*n* = 47) [GSE28572]**	**Short-Term Survivors (n =53) [GSE2852]**
Gender				
Male	-	-	24	30
Female	-	-	23	23
Race				
White	-	-	NA	NA
Black or African-American	-	-	NA	NA
Asian	-	-	NA	NA
NA	-	-	NA	NA
Histological Type				
Lung adenocarcinoma	179	-	24	26
Lung squamous cell carcinoma	92	-	15	13
Lung large-cell carcinoma	None	-	8	14
Age				
Age, median (range), yr	-	-	63	66
Survival, median (range), mo	-	-	107	9
Vital Status				
Alive	-	-	NA	NA
Dead	-	-	NA	NA
Tumor State				
IA–IIB	-	-	38	40
IIIA–IV	-	-	6	12
Missing	--	-	3	1
Ethnicity				
Not Hispanic or Latino	-	-	NA	NA
Hispanic or Latino	-	-	NA	NA
NA		-	NA	NA
Tobacco Smoking History				
Never smoker	-	-	3	4
Ex-smoker	-	-	20	22
Current smoker	-	-	22	25
Missing	-	-	2	2

3.7 Obtaining Discretized DNA CNV Calls and Gene-Probe Matching

The usage of three different datasets required employment of different methods of data processing, before Boolean implication relationships could be detected in a cohort. However, the sequential flow of the procedure used can be generalized in the following steps:

- Processing sample data to obtain DNA CNV calls for each sample in every probe.
- Converting probe names to gene names in order to examine the interaction amongst genes.
- Running the Genet-CNV algorithm on processed data to finally obtain Boolean implication interactions

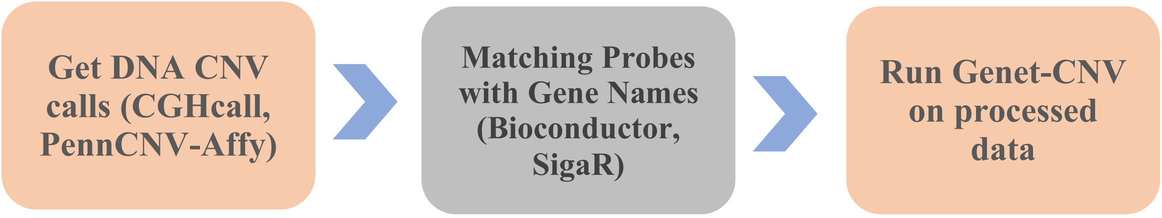

Get DNA CNV calls (CGHcall, PennCNV-Affy)
Matching Probes with Gene Names (Bioconductor, SigaR)
Run Genet-CNV on processed data

Figure 3.6 Workflow for Obtaining Input for Gene-CNV from Raw CNV microarray data

3.7.1 GSE31800

Dataset GSE31800 comprises aCGH data with log2 normalized ratios of Cy3/Cy5. For this dataset, CGHcall (van de Wiel et al., 2007) package was implemented in R in order to obtain discrete calls for CNV for each instance in the dataset. CGHcall is a software tool to classify aCGH data into copy number states (deletion, loss, normal, gain or amplification). CGHcall has certain advantages over other tools with similar functionality. It uses DNACopy (Venkatraman & Olshen, 2007), a circular binary segmentation algorithm, which has been shown to be one of the strongest segmentation algorithms for CNV analysis. CGHcall also analyzes for six states, instead of the usual three states of gain, loss and normal that most segmentation algorithms do. The six states reflect double deletion, single deletion, normal, gain, double gain and amplification.

For the first dataset, DNA copy number profiles of 271 samples, with log2 normalized ratios given across 53,856 probes were provided as input to CGHcall, in order to yield the number of calls as the processed output. The output resulted in 49,710 probes for the 271 samples in the dataset. Approximately 4000 of the probes were excluded from CNV calls evaluation due to missing data or errors arising during pre-processing.

This was followed by the process of translation of probe IDs to obtain gene names. This dataset had DNA copy number data and gene expression profile on the same microarray platform, thus enabling the usage of gene probes as mentioned in the gene expression data to obtain gene names for the probes used in DNA CNV data generation.

MatchMiner was used in Batch Lookup mode to acquire the corresponding cytogenetic locations of the list of genes used in the gene expression analysis. The list of probe names from the RNA samples were converted by MatchMiner to Gene symbols. A list of chromosome base pair starting and ending positions, along with their corresponding cytogenetic bands for hg19, the human genome references as used by the UCSC Genome Browser, was obtained. The list of gene symbols obtained from MatchMiner was matched with its corresponding list of cytogenetic bands over NCBI database. Thus, using cytogenetic bands as the common factor, gene symbols were matched with chromosome base-pair starting and ending positions. This final list of gene symbols with their chromosome base-pair starting and ending positions was processed along with the chromosome base-pair locations provided in GSE31800 DNA CNV samples, using the R package sigaR. This package is based on the Distance Matching algorithm (van Wieringen, Belien, Vosse, Achame, & Ylstra, 2006; van Wieringen et al., 2012). Thus, we had gene symbols for most of the probes in GSE31800 DNA CNV data. Most gene symbols corresponded to two sets of probes in the dataset, and in such cases, only the probe set with the higher value of CNV calls was included in the final analysis. This led to the final matrix of DNA CNV calls with 271 samples and 19,720 genes, to be analyzed using the Boolean Implication algorithm.

3.7.2 GSE71294

For the second dataset, CGHcall was used for obtaining the DNA CNV calls. The procedure was similar to that followed for the second database. For this dataset, the gene name corresponding with each probe was provided in the microarray platform information [GPL4091]. After removing all the repeated probe names, and any probes with missing data removed during CNV analysis, the final matrix was created. It comprised 16,122 unique genes, for 64 given patient samples.

3.7.3 GSE28572

The third dataset is based off an SNP array and as such required a different package and methodology for processing for DNA CNV calls. PennCNV-Affy (Wang et al., 2007), which is a software for Copy Number Variation (CNV) detection from SNP genotyping arrays, was used to obtain CNV calls. Instead of using a segmentation-based algorithm, it uses a Hidden Markov Model (HMM), to infer CNV calls (Wang et al., 2007).

The raw data in the form of CEL files available publicly was required for this analysis. Affymetrix Power Tools (Scientific, 2019, January 13) application was used to process the CEL files and gather the normalized signal intensity data. The PennCNV-Affy tool is then applied to obtain the final calls. The output is in the form of a table that lists all the CNVs discovered. The table has fields that list how many SNPs are contained within each CNV and the length of the CNV. There is also a field to identify the CN state of the CNV, where CN < 2 is classified as a deletion, whereas CN > 2 means there is a duplication. A program was written to integrate this data into a matrix with the CN state of for each probe and each sample.

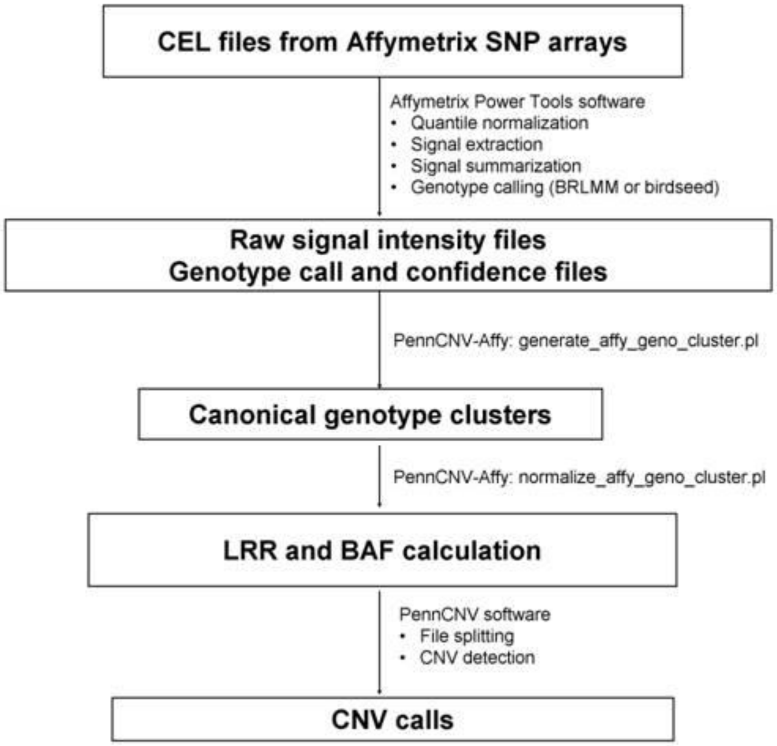

CEL files from Affymetrix SNP arrays
Affymetrix Power Tools software
• Quantile normalization
• Signal extraction
• Signal summarization
• Genotype calling (BRLMM or birdseed)
Raw signal intensity files
Genotype call and confidence files
PennCNV-Affy: generate_affy_geno_cluster.pl
Canonical genotype clusters
PennCNV-Affy: normalize_affy_geno_cluster.pl
LRR and BAF calculation
PennCNV software
• File splitting
• CNV detection
CNV calls

Figure 3.7 Workflow for the Application of PennCNV-Affy Algorithm to SNP microarray data (PennCNV, 2019)

The SNP probes were converted into gene names by using rsIDs given in the microarray platform data, and converting them into gene IDs using 'ensembledb' library in 'Bioclite' package in R. Since an SNP represents one nucleotide, multiple consecutive SNPs translated into the same gene names. For most of these SNPs falling within the same gene, the distribution of CNVs was the same. In cases where such was not the case, the probes with the maximum intensity were selected.

This matrix was further broken down into sub sets of 47 samples (long survival) and 53 samples (short survival) for differential analysis of genes in long and short survival patient samples.

3.8 Analysis with Genet-CNV

After generating CNV calls, Genet-CNV was used to discover pairwise gene-association rules from the three patient cohorts. All the results obtained were at a 95% significance level, unless mentioned otherwise.

3.8.1 Exclusion of Rules with Precision Levels Close to Zero

In order to carry out one-tailed Z-test based on the sample size, the mean value of precision and scope for all possible rules from a given a dataset is calculated.

A remarkable observation was made while calculating the mean precision value in all three datasets. It was noticed that the value of average precision tended to be less than 0.10 when no rules were excluded when averaging the precision. When the 5% rules with the lowest precision values were excluded, the average precision value rose up to approximately 0.50. The trend continued when the 10% rules with the lowest precision values were excluded while calculating mean precision. The average precision value rose up slightly in this case. However, when all rules were included, the average precision value showed a sheer drop to around 0.10.

We speculate that the reason for the average precision of the rules coming out to be so low when all the rules are included could be because of the housekeeping genes present in the datasets. These genes are not actively involved in any interactions, and therefore lower the mean precision value for all the genes present in the dataset. However, when at least the bottom-most 5% of the rules are filtered out, most of these housekeeping genes must be getting filtered out and the average precision value rises up to around 0.50.

In the table below, the first column denotes the percentage of rules with the minimum precision values that were excluded for calculating the mean precision for Co-amplification and Co-deletion interactions in each of the three datasets. The other cells in the table display the average precision obtained at each threshold.

Table 3.2 Mean Precision at Different Thresholds for Exclusion

Threshold for Exclusion	Mean Precision					
	GSE31800		GSE72194		GSE28572	
	Amp	Del	Amp	Del	Amp	Del
0%	0.071682	0.082772	0.137554	0.244195	0.062087	0.082722
5%	0.599896	0.553061	0.702721	0.588192	0.747988	0.553061
10%	0.66374	0.615929	0.742394	0.642396	0.778604	0.615929

We continued calculating the mean precision value for different thresholds of exclusion at intervals of 10%, and plotted the curve obtained for each of the four cases: co-amplification, co-deletion, amplification-deletion and deletion-amplification. Following is the curve obtained for GSE72194:

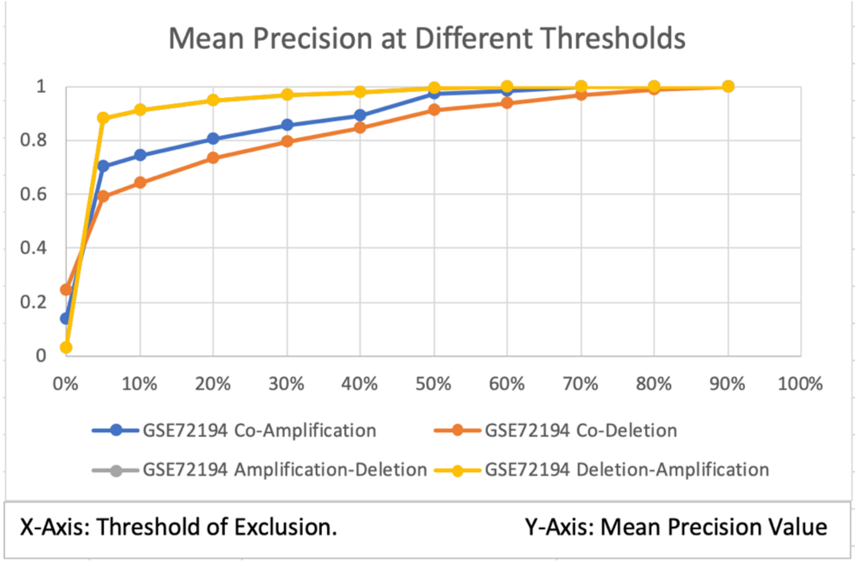


Fig 3.8 Mean Precision at Different Threshold of Exclusion [GSE72194]

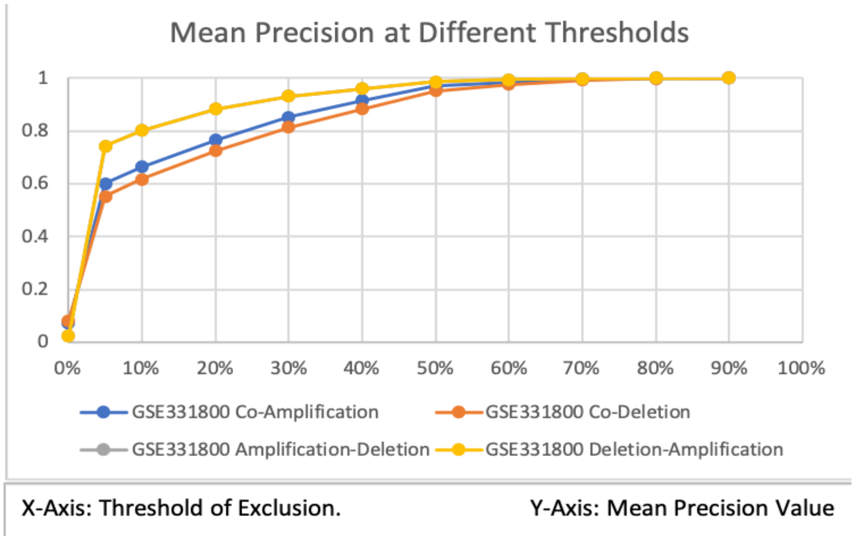


Fig 3.9 Mean Precision at Different Threshold of Exclusion [GSE31800]

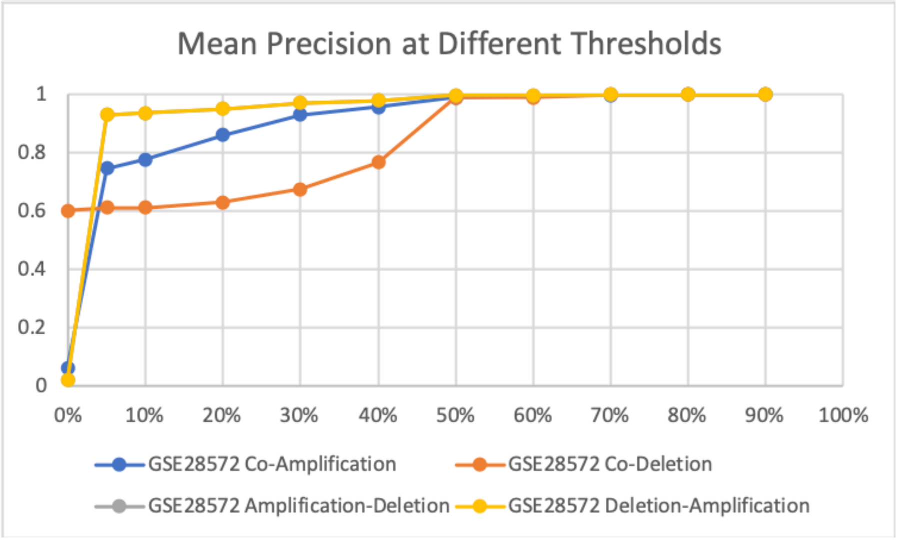


Fig 3.10 Mean Precision at Different Threshold of Exclusion [GSE28572]

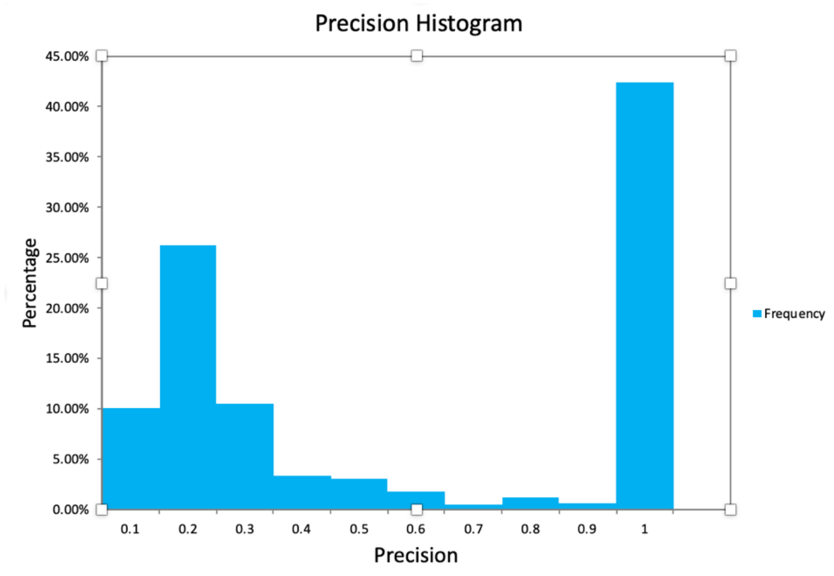


Fig 3.11 Precision Histogram [GSE31800, co-amplification]

It was observed that the precision value with the maximum frequency was 1.0, for the three patient cohorts. A precision of 1 is calculated when the number of error cells is equal to zero. In DNA CNV data, the number of samples that show a variation from normal copy number (1 for amplified and -1 for deleted) are very small. Therefore, it is quite likely that the number of error cells, and hence the precision calculates to zero. This is explained with the assistance of an example rule that was observed in GSE31800 co-amplification rules.

The gene pair CD27-MBD2 were evaluated to be of the negative implication rules, i.e. rule type 4. The following is a contingency table for the given rule:

Table 3.3 Contingency Table for CD27-MBD2

A⇒B	***A⇒¬B***
5	6
¬A⇒B	***¬A⇒¬B***
0	260

Since this was evaluated as rule type 4, therefore the number of error cells in this case would be the frequency for A'B, which is 0 in this case. According to the formula for precision for rule 4, precision is 1 if the number of error cells is zero:

$$\nabla_{\neg A\Rightarrow\neg B} = \nabla_{\neg A^\wedge B} = 1 - \frac{N_{\neg A^\wedge B}}{N * U_{\neg A\Rightarrow\neg B}}$$

Therefore, the precision is given as 1.

With our sparse data, it is very likely to get the number of error cells as 0, and hence perfect precision, and Genet-CNV picks the rule with the highest precision in all cases.

3.8.2 Logical Equivalence of Rules

In Boolean logic, if we have a proposition such as A⇒ B, then applying modus tollens on this, we also have ¬B ⇒ ¬ A. The following table summarizes the logical equivalence of the four rules used in Boolean implication algorithm, when modus tollens is applied to these rules:

Table 3.4: Logical Equivalence of Rules when Modus Tollens is Applied

Rule	**Logical Equivalence using Modus Tollens**
A⇒B	¬B ⇒ ¬A
A⇒¬B	B ⇒ ¬A
¬A⇒B	¬B ⇒ A
¬A ⇒ ¬ B	B ⇒ A

In figures 3.8, 3.9, and 3.10, It is observed that there are only three different curves visible in the graph, even though we are plotting four sets of interactions: co-amplification, co-deletion, amplification-deletion, and deletion-amplification.

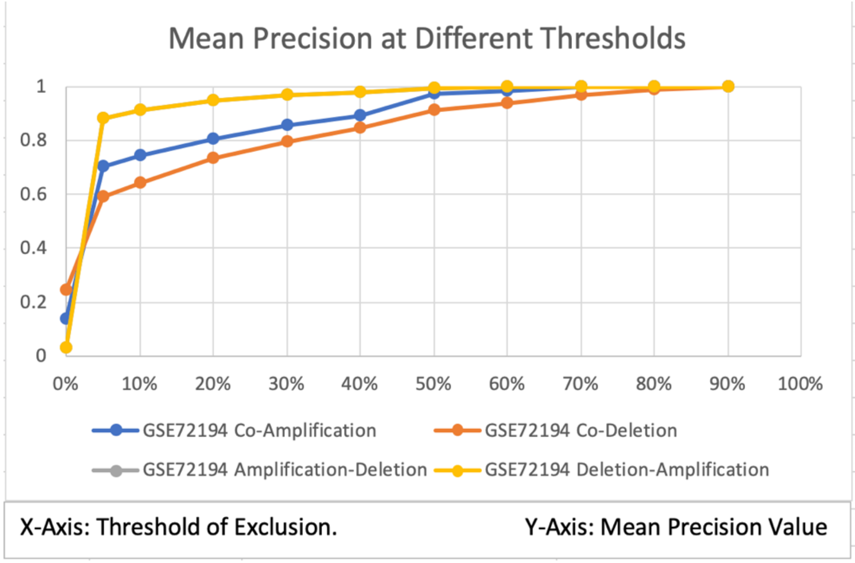


Fig 3.12: Only three of the four series of data are visible

Using Genet-CNV on DNA CNV dataset, we notice that the rules for amplification-deletion and for deletion-amplification are logical equivalents of each under when applying modus tollens. Thus, we get the same set of rules for both amplification-deletion, and deletion-amplification, even though they fall under different rule types in the two interactions. The following table details the rules as applicable to our datasets, and their logical equivalents using modus tollens.

Table 3.5: Logical Equivalence Using Modus tollens for Amplification-Deletion and Deletion-Amplification

Interaction Type	Rule Type	Total Rules			
		Rule 1	**Rule 2**	**Rule 3**	**Rule 4**
Amplification -Deletion	Rules	1 & -1	1 & !-1	!1 & -1	!1 & !-1
	Logical Equivalents	!-1 & !1	-1 & !1	!-1 & 1	-1 & 1
Deletion -Amplification	Rules	-1 & 1	-1 & !1	!-1 & 1	!-1 & !1
	Logical Equivalents	!1 & !-1	1 & !-1	!1 & -1	1 & -1

From the table it is clear that the logical equivalent for rule 1 for amplification-deletion is the same as rule 4 for deletion-amplification, and the logical equivalent for rule 1 for deletion-amplification is the same as the rule 4 for amplification-deletion. The logical equivalent for rule 2 for amplification-deletion is the same as rule 2 for deletion-amplification, and the logical equivalent of rule 3 in one interaction corresponds to rule 3 in the other.

Thus, the same set of total rules is obtained for both these interaction types, and therefore, the curve for mean precision is precisely the same at different thresholds for amplification-deletion and deletion-amplification. The curves in the figures therefore get superimposed upon one another.

3.8.3 Biological Significance of Rules 1 and 5

Rule 1 and Rule 5 obtained by Genet-CNV, which stand for positive implication and positive equivalence respectively, are rich in biological information. Depending upon the case in question, they describe co-amplification, co-deletion, amplification-deletion and deletion-amplification relationships, etc. Looking at co-amplification for example, if rule 1 is found to be true for a pair of rules, gene A and gene B, then this implies that if gene A is amplified, then gene B is also amplified. If rule 5 is found for the same pair, then this implies that if gene A is amplified, then gene B is amplified, and also if gene A is not amplified, then gene B is not amplified. Both these rules have straightforward biological interpretations and are of significant interest when looking at genes in network analysis. Rule four which stands for negative implications, is not of significant biological interest, as it is looking at a non-occurrence event. If rule 4 is found to be true for a pair of genes, when looking at co-amplification once again, that this means if gene A is not amplified, then gene B is not amplified. Events where both the genes are not amplified are not of particular note to us, since it gives no information about amplification.

3.8.4 Comparison of Total Number of Rules at Different Significance Levels

We then examined the total number of rules obtained at different significance levels. As expected, decreasing the significance yields a much larger number of rules and vice versa. The following table gives the number of rules obtained at different significance levels for the network of seven genes with the genome:

Table 3.6 Total number of rules obtained [seven genes] for three different significance levels

Datasets	Total Rules		
	p-value = 0.025	p-value = 0.05	p-value = 0.10
GSE31800	5108	12091	20522
GSE72194	13288	23975	29573
GSE28572	383	3074	5871

3.9 Data Analysis Using Python over Spruce Knob Cluster for High Performance Computing

A very large number of rules were identified using Genet-CNV, with the largest number of rules being 33 million for whole genome interaction in GSE31800. In order to compile lists of rules common amongst the datasets, and in further evaluation of results obtained by Genet-CNV with the MSigDB database, scripts were written in Python to be run over Spruce Knob HPC clusters. **Pandas** and **Numpy** libraries in Python were used for sorting through the gene interaction tables and identifying common rules and in evaluation with MSigDB.

Code written in python was run on Spruce Knob High Performance Computing Clusters. Compute resources for one cluster node with 6 processors and 16 GB memory per processor were utilized to run the code. It took 4 minutes for the largest datasets to be compared. This same comparison could not be replicated on macOS Mojave with a 2.3GHz Intel Core i5 processor and 8 GB 2133 MHz LPDDR3 RAM. Not enough memory resources were available to hold the datasets in the memory and perform computations on them.

Chapter 4

Results and Evaluation

4.1 Rules for Seven Genes and Whole Genome

Genet-CNV was used to generate implication rules between gene pairs for genome-wide data. The result was a large number of rules which could be mined and analyzed to discover pertinent ones and further validate them using biological experiments. The datasets under scrutiny were analyzed to generate rules with two separate outcomes in mind. In the first approach, the interaction of seven specifically identified genes with the rest of the genome was observed to generate implication rules between each of these seven genes and the genes in the rest of the genome. In the second approach, all the genomes present in a cohort were compared with every other gene present in the same cohort. This second approach resulted in the generation of a huge number of rules, sometimes numbering in the millions. This was further used to compare genetic interactions common across all three of the datasets.

4.1.1 Selection of Seven Genes that Serve as Prognostic Biomarkers for NSCLC

The first approach, i.e. the examination of seven genes with the rest of the genome, was based on a study conducted by Guo et al. This study presented "a predictive multi-gene assay and prognostic protein biomarkers clinically applicable for improving NSCLC treatment, with important implications in lung cancer chemotherapy and immunotherapy" (N. L. Guo et al., 2018). The genome-wide transcriptional profiles and qRT-PCR were used to generate a multi-gene assay, for predicting the prognosis of NSCLC cases, and benefits of chemotherapy. This multi-gene assay was further validated by examining protein cohorts from independent data. The protein expression of the seven genes in this assay was correlated with the mRNA expression and DNA copy number variation from patient tissue samples for validation of functional involvement and potential as therapeutic targets in chemotherapy and immunotherapy.

The prognostic biomarkers used in this study were evaluated using Cox proportional hazard model. The hazard ratio of each biomarker was calculated. A hazard ratio of greater than 1 meant that the gene under scrutiny is associated with poor outcome when down-regulated, but up-regulation of the same gene is associated with a good outcome in NSCLC patients. A hazard ratio of less than 1 implied that down-regulation is associated with good outcome, and up-regulation with poor outcome. Finally, seven genes were selected as prognostic classifier based on decision trees. These seven genes are: **ABCC4, CCL19, SLC39A8, CD27, FUT7, DAG1 and ZNF71.** This seven-gene prognostic model was further validated on independent patient cohorts.
Considering that these seven genes have been identified as prognostic biomarkers in the case of NSCLC,

their interactions with other genes in the genome could help prepare hypothesis for further investigation of biologically significant genes and interactions in NSCLC cases.

In the GSE31800 dataset, all seven of these prognostic biomarkers were present. In GSE72194, 6 of the seven genes were present, FUT7 was absent. In GSE28572, three out the seven genes, ABCC4, ZNF71 and SLC39A8 were found. The following tables display the number of DNA CNV found for each gene in each of the three patient cohorts:

Table 4.1 CNV States of Seven Genes in the Three Patient Cohorts (Seven Genes)

Gene Name	GSE31800 [n = 271]			GSE72194 [n= 64]			GSE28572 [n = 100]		
	Gain	Loss	Normal	Gain	Loss	Normal	Gain	Loss	Normal
CD27	11	2	258	4	7	53	-	-	-
FUT7	2	51	218	-	-	-	-	-	-
ZNF71	5	19	247	3	4	57	7	4	89
DAG1	0	20	251	0	19	45	-	-	-
ABCC4	3	3	265	3	14	47	5	4	91
CCL19	1	34	236	4	4	56	-	-	-
SLC39A8	3	4	264	0	0	64	0	1	99

Table 4.2 CNV States of Seven Genes for Long and Short Survival patient samples in GSE28572

Gene Name	GSE28572, Long Survival [n = 47]			GSE28572, short Survival [n = 53]		
	Gain	Loss	Normal	Gain	Loss	Normal
CD27	-	-	-	-	-	-
FUT7	-	-	-	-	-	-
ZNF71	1	1	45	6	3	44
DAG1	-	-	-	-	-	-
ABCC4	4	1	42	1	3	49
CCL19	-	-	-	-	-	-
SLC39A8	0	0	47	0	1	52

4.1.2 Whole Genome Comparison

The second approach was to compare the whole genome with the whole genome. This led to the discovery of a significantly larger number of rules than in the prior case, as the number of comparisons involved was so much larger.

4.2 All Rules Obtained with Genet-CNV

4.2.1 Rules from Three Patient Cohorts

The following two tables list the total number of rules obtained for the three patient cohorts, firstly with the seven genes previously identified, followed by whole-genome interactions. The significance level for all following results henceforth, is 95% (p < 0.05), unless specified otherwise. The number of rules for each rule type in each patient cohort is listed in Appendix A.

Table 4.3 Number of rules for each interaction type in each data set (Seven Genes)

Type of Interaction	Datasets		
	GSE31800	GSE72194	GSE28572
Co-Amplification	1524	2816	324
Co-Deletion	7320	20506	2492
Amplification-Deletion	514	97	113
Deletion-Amplification	2733	556	145
Total	12091	23975	3074

Table 4.4 Rule Types for GSE31800 (Seven Genes)

GSE31800	Rule 1	Rule 2	Rule 3	Rule 4	Rule 5	Rule 6	Total
Co-Amplification	1393	0	0	60	71	0	1524
Co-Deletion	3030	169	0	3296	825	0	7320
Amplification-Deletion	484	0	0	30	0	0	514
Deletion-Amplification	659	1962	0	112	0	0	2733

Table 4.5 Rule Types for GSE72194 (Seven Genes)

GSE72194	Rule 1	Rule 2	Rule 3	Rule 4	Rule 5	Rule 6	Total
Co-Amplification	2322	0	0	397	97	0	2816
Co-Deletion	7656	0	0	9823	3027	0	20506
Amplification-Deletion	38	0	0	0	59	0	97
Deletion-Amplification	20	413	0	123	0	0	556

Table 4.6 Rule Types for GSE28572 (Seven Genes)

GSE28572	Rule 1	Rule 2	Rule 3	Rule 4	Rule 5	Rule 6	Total
Co-Amplification	148	0	0	172	4	0	324
Co-Deletion	64	0	0	2106	322	0	2492
Amplification-Deletion	0	0	0	113	0	0	113
Deletion-Amplification	110	0	0	35	0	0	145

Table 4.7 Number of rules for each interaction type in each data set (Whole Genome)

Type of Interaction	Datasets		
	GSE31800	**GSE72194**	**GSE28572**
Co-Amplification	12184526	2605585	3353315
Co-Deletion	14443017	26584060	4090758
Amplification-Deletion	4647520	471845	384439
Deletion-Amplification	4647520	471845	384439
Total	35922583	30133335	8212951

Table 4.8 Rule Types for GSE31800 (Whole Genome)

GSE31800	Rule1	Rule2	Rule3	Rule4	Rule5	Rule6	Total
Co-Amplification	6456569	294304	0	4999054	434599	0	12184526
Co-Deletion	6077814	593030	0	6077814	1694359	0	14443017
Amplification-Deletion	670347	3561985	0	414528	660	0	4647520
Deletion-Amplification	414528	3561985	0	670347	660	0	4647520

Table 4.9 Rule Types for GSE72194 (Whole Genome)

GSE72194	Rule1	Rule2	Rule3	Rule4	Rule5	Rule6	Total
Co-Amplification	1551716	742	0	873760	179367	0	2605585
Co-Deletion	10936327	1540	0	10936327	4709866	0	26584060
Amplification-Deletion	45100	306004	0	73691	47050	0	471845
Deletion-Amplification	73691	306004	0	45100	47050	0	471845

Table 4.10 Rule Types for GSE28572 (Whole Genome)

GSE28572	Rule1	Rule2	Rule3	Rule4	Rule5	Rule6	Total
Co-Amplification	1822103	3316	0	1341863	186033	0	3353315
Co-Deletion	999126	0	0	999126	2092506	0	4090758
Amplification-Deletion	28162	0	0	344235	12042	0	384439
Deletion-Amplification	344235	0	0	28162	12042	0	384439

All the rules found for the seven genes were also present amongst the rules discovered when comparing the whole genome. Therefore, it can be said that the rules in Table 4.3 are a subset of the rules in Table 4.7.

4.2.2 Rules from Long-Term and Short-Term Survival Patient Samples [GSE28572]

As mentioned previously, for one of the datasets, GSE28572, the patient samples were classified into two categories: long survival and short survival. Looking at both, genes that are common, and genes that are unique in the two categories, could be used to further investigate the correlation of certain genes with survival status in NSCLC patients. As a result, we ran two analyses on GSE28572, running Genet-CNV on long survival patients and short survival patients separately. We then compared the rules found in short survival patient samples with those found in long survival patient samples. The tables below give the number of rules obtained for each interaction type, in the case of the seven genes, followed by that of the whole genome.

Table 4.11 Number of rules for each interaction type in Survival Cohort (Seven Genes)

Type of Interaction	GSE28572	
	Long Survival	Short Survival
Co-Amplification	137	133
Co-Deletion	0	317
Amplification-Deletion	0	0
Deletion-Amplification	0	9
Total	137	459

Table 4.12 Rule Types for Long Survival, GSE28572 (Seven Genes)

Long Survival, GSE28572	Rule 1	Rule 2	Rule 3	Rule 4	Rule 5	Rule 6	Total
Co-Amplification	121	3	0	0	16	0	137
Co-Deletion	0	0	0	0	0	0	0
Amplification-Deletion	0	0	0	0	0	0	0
Deletion-Amplification	0	0	0	0	0	0	0

Table 4.13 Rule Types for Short Survival, GSE28572 (Seven Genes)

Short Survival, GSE28572	Rule 1	Rule 2	Rule 3	Rule 4	Rule 5	Rule 6	Total
Co-Amplification	61	0	0	72	0	0	133
Co-Deletion	9	0	0	1	308	0	317
Amplification-Deletion	0	0	0	0	0	0	0
Deletion-Amplification	1	0	0	0	8	0	9

Table 4.14 Number of rules for each interaction type in Survival Cohort (Whole Genome)

Type of Interaction	GSE28572	
	Long Survival	Short Survival
Co-Amplification	1003363	888696
Co-Deletion	481662	4732
Amplification-Deletion	8939	7324
Deletion-Amplification	8939	7324
Total	1502903	908076

Table 4.15 Rule Types for Long Survival, GSE28572 (Seven Genes)

Long Survival, GSE28572	Rule1	Rule2	Rule3	Rule4	Rule5	Rule6	Total
Amplification	530157	6	0	259982	98551	0	888696
Deletion	37	0	0	37	4658	0	4732
Amplification-Deletion	95	0	0	6596	633		7324
Deletion-Amplification	6596	0	0	95	633	0	7324

Table 4.16 Rule Types for Long Survival, GSE28572 (Seven Genes)

Short Survival, GSE28572	Rule1	Rule2	Rule3	Rule4	Rule5	Rule6	Total
Amplification	593956	56	0	307716	101635	0	1003363
Deletion	5831	0	0	5831	470000	0	481662
Amplification-Deletion	1163	0	0	4931	2845	0	8939
Deletion-Amplification	4931	0	0	1163	2845	0	8939

4.3 Identification of Common Rules

In order to identify gene interactions with a strong correlation, the rules generated by Genet-CNV were mined to detect the ones that were common in any two databases or common across all. The following tables list the number of common rules found amongst all three datasets, and the number of common rules between long term survival samples and short-term survival samples for the GSE28572 Cohort.

4.3.1 Identification of Common Rules for three Patient Cohorts

Table 4.17 Number of Common Rules (Seven Genes with Whole Genome)

Type of Interaction	Datasets			
	GSE31800 & GSE72194	GSE72194 & GSE28572	GSE28572 & GSE31800	All Datasets
Co-Amplification	52	50	28	20
Co-Deletion	294	81	20	0
Amplification-Deletion	0	0	0	0
Deletion-Amplification	0	0	1	0
Total	346	131	49	20

Table 4.18 Number of Common Rules (Whole Genome with Whole Genome)

Type of Interaction	Datasets			
	GSE31800 & GSE72194	GSE72194 & GSE28572	GSE28572 & GSE31800	All Datasets
Co-Amplification	294032	85835	96294	15641
Co-Deletion	1080948	20146	23040	2572
Amplification-Deletion	6606	21	292	0
Deletion-Amplification	6606	21	292	0
Total	1388192	106023	119918	18213

4.3.2 Identification of Common Rules for Long and Short Survival [GSE28572]

There were no common rules found between the set of long-term survival patient samples and short-term survival patients in patient cohort GSE28572. The following graph visualizes the network of gene interactions discovered for long-term survival in GSE28572. There were 137 rules discovered, with 121 rules denoting co-amplification between ABCC4 and other genes, and 16 equivalence rules between ABCC4 and other genes.

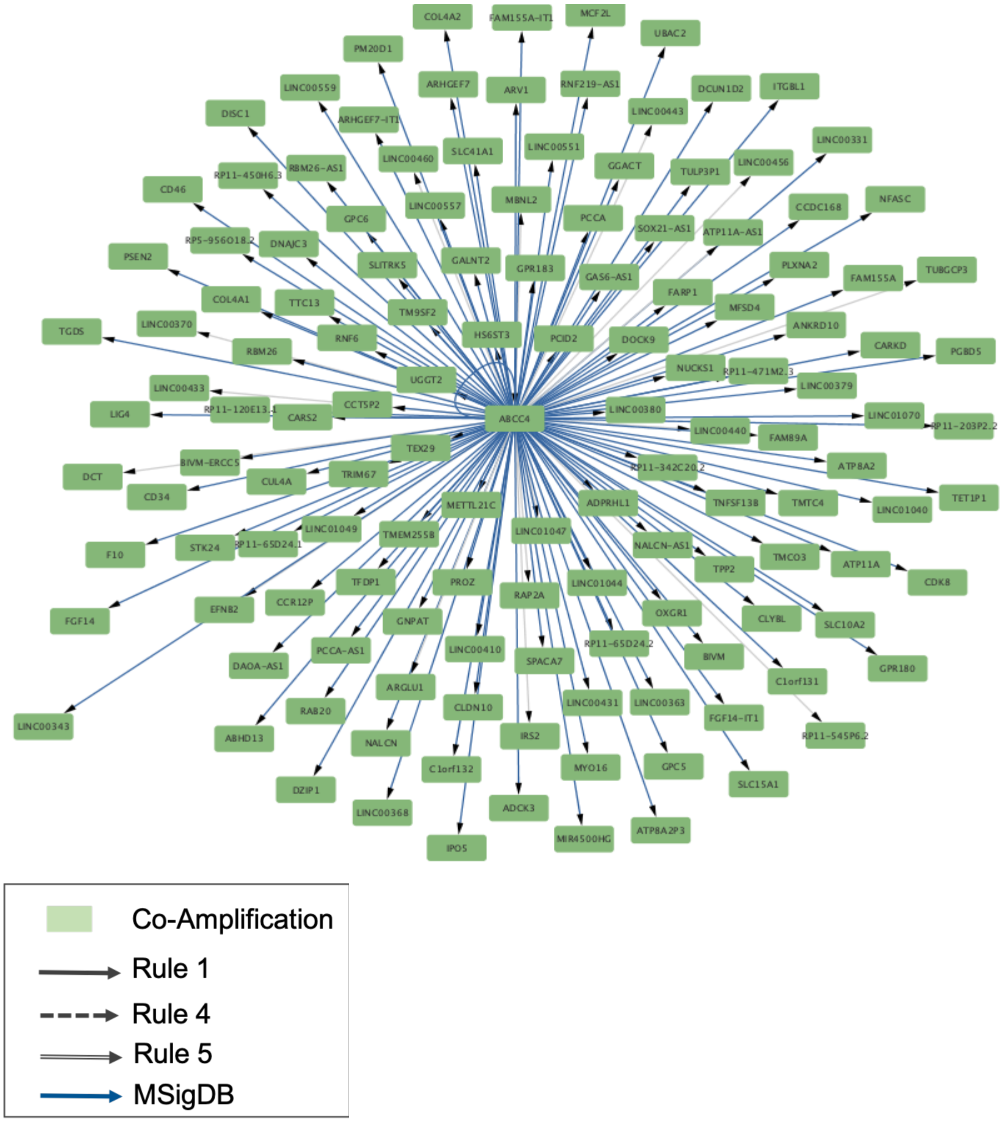


Figure 4.1 Long Survival [GSE28572]: All Rules

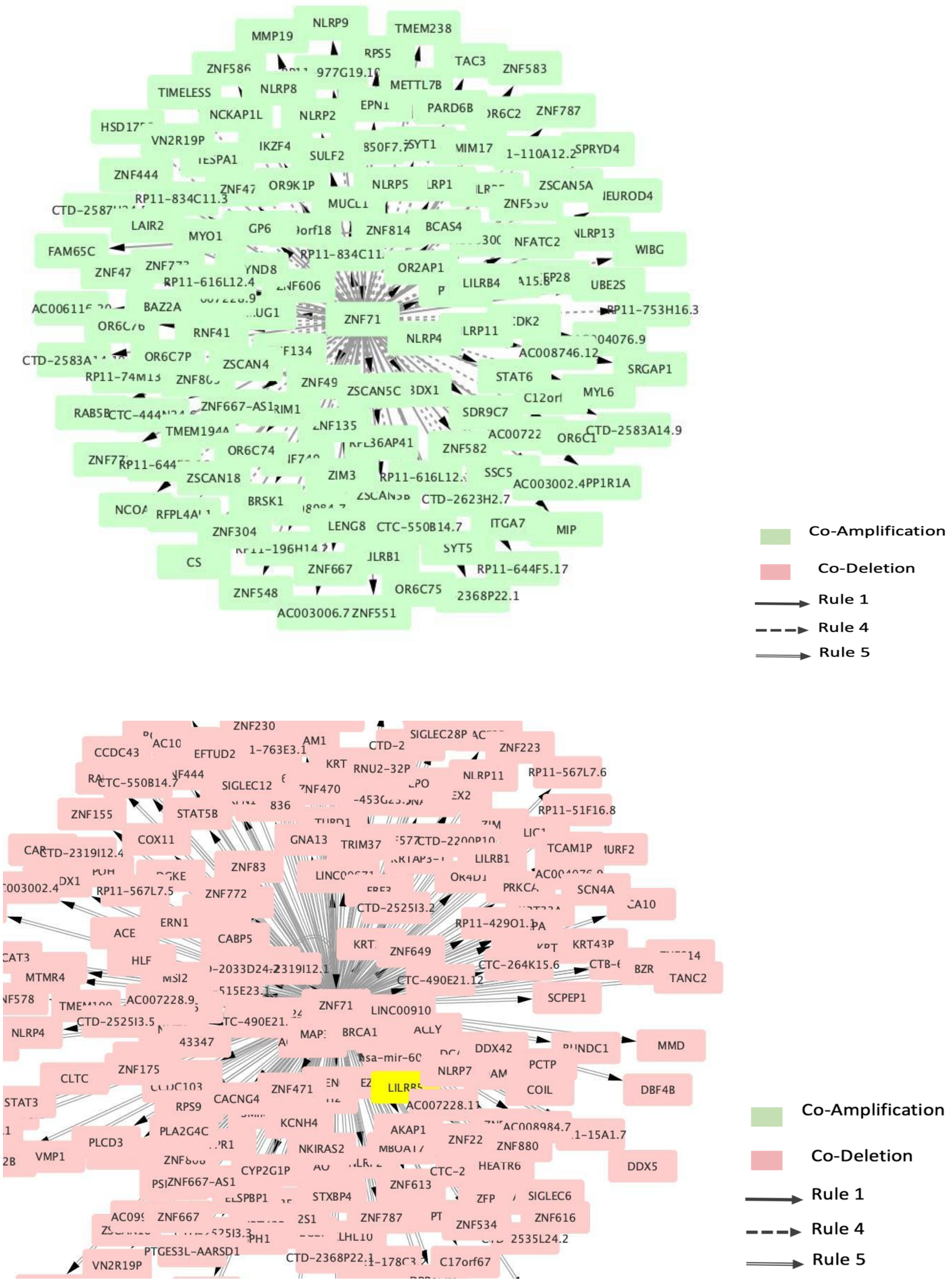


Figure 4.2 Short Survival [GSE28572]: Amplification and Deletion

4.4 Comparison with MSigDB

The implication rules thus obtained from the three datasets after using Genet-CNV, were further compared with the Gene Sets in the Molecular Signature Database (MSigDB) (Liberzon et al., 2015; Liberzon et al., 2011). The Molecular Signatures Database (MSigDB) is a collection of annotated gene sets for use with GSEA software. The datasets in MSigDB detail the gene interaction pathways that have been compiled from studies and other published works, and as such, is the benchmark dataset for gene associations. MSigDB is organized into eight major gene sets, which include the following:

- Hallmark Gene Sets (H)
- Positional Gene Sets (C1)
- Curated Gene Sets (C2)
- Motif Gene Sets (C3)
- Computational Gene Sets (C4)
- GO Gene Sets (C5)
- Oncogenic Signatures (C6)
- Immunologic Signatures (C7)

Each of these gene set further comprises several sub-collections.

We compared all the Boolean implication rules yielded by the application of Genet-CNV on each dataset with the all the sub-collections present in all the right gene sets in MSigDB. The following table lists the number of interactions found to be common between the gene interactions from our three datasets and the genes present in the MSigDB gene sets.

Table 4.19: Number of Matches Found Between Genet-CNV Results and MSigDB Results for the Seven Genes with Whole Genome

Datasets	Total Unique Rules Obtained by Genet-CNV	Unique Rules Present in MSigDB
GSE31800	11176	6283
GSE72194	22981	11703
GSE28572	2954	1055
GSE28572 (Long Survival)	139	83
GSE28572 (Short Survival)	459	137

For the interactions amongst seven genes and the rest of the genome, out of the 20 rules common in all three datasets, 18 were validated with MSigDB as well. The following figure describes the network.

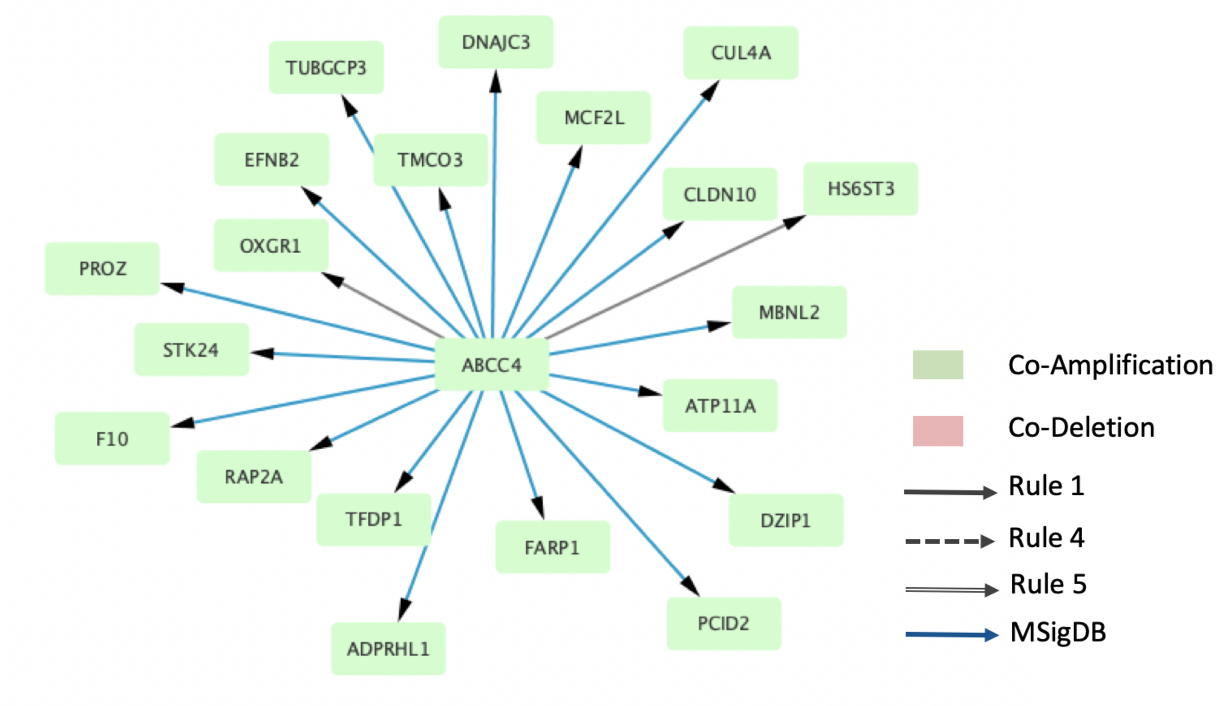


Figure 4.3 Common Rules for Seven Genes Validated with MSigDB

The number of comparisons for the whole genome results with MSigDB are not listed, as tens of millions of rules and to compare them with MSigDB would be beyond the scope of this study.

We did, however, compare the set of rules common to all three datasets in whole genome comparison, and validated them against MSigDB. 15,436 out of the 18,213 common rules identified were validated with MSigDB.

4.5 Proliferation, Apoptosis, Survival and Immunologic Pathways

The set of rules common in all three datasets for whole genome comparisons was matched with MSigDB. 15,436 out of the 18,213 common rules identified were validated with MSigDB. From these common and validated rules, we identified gene associations that were present in four specific pathways which are of special interest in understanding the disease mechanisms of NSCLC. These four pathways included apoptosis, proliferation, survival and immunologic pathways.

Apoptosis, or programmed cell-death, occurs as a normal and controlled part of a cell's life cycle. Interference or malfunction in this process plays an instrumental role in cancer. Proliferation, survival, and immunologic processes are all important factors to be investigated when it comes to cause and treatment of cancer. Thus, the gene associations thus identified for these pathways, which were present in all three datasets and were also validated with MSigDB, could be very useful in generating insights for biologists to validate.

The following table lists the number of genetic interactions identified for each of the four pathways:

Table 4.20 Number of Common Rules Present in the Four Pathways

Pathways	Unique Rules	Unique Genes
Survival	78	81
Apoptosis	459	319
Proliferation	412	194
Immunology [C7]	7262	1652

The following figure depicts the network of interactions for the survival pathways.

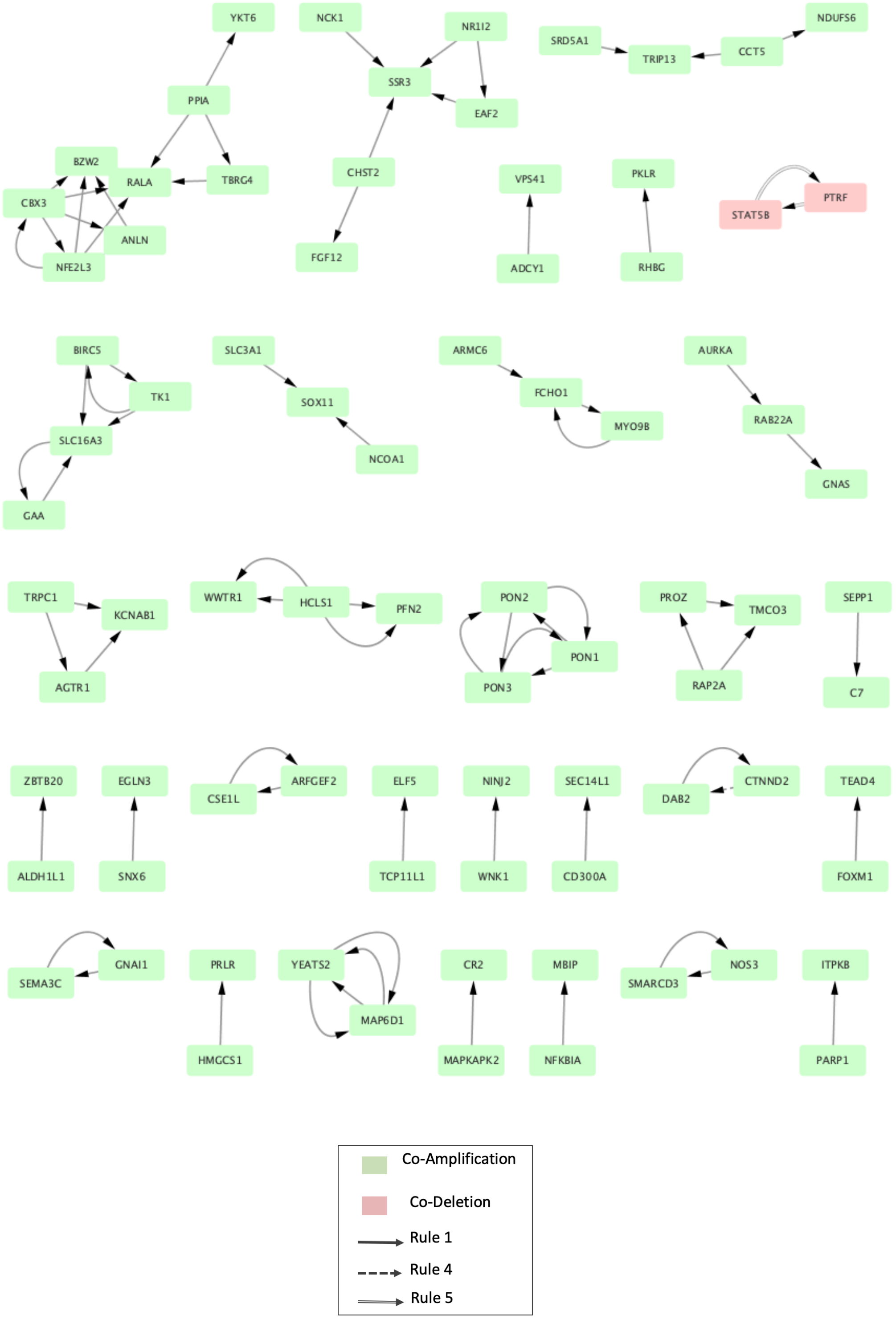


Figure 4.4 Common Rules in the MSigDB Survival Pathways

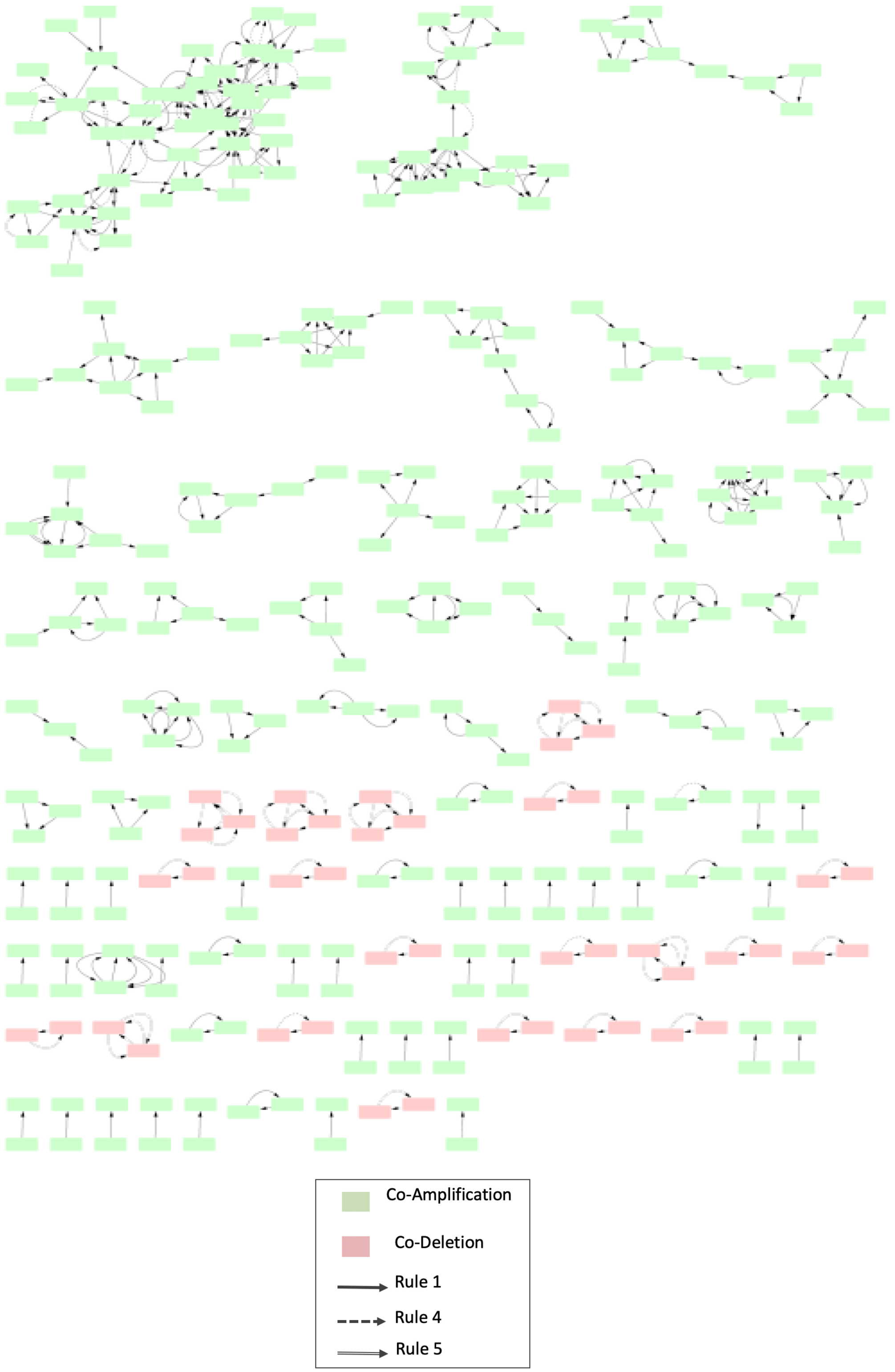


Figure 4.5 Representation of the Common Rules in the MSigDB Apoptosis Pathways

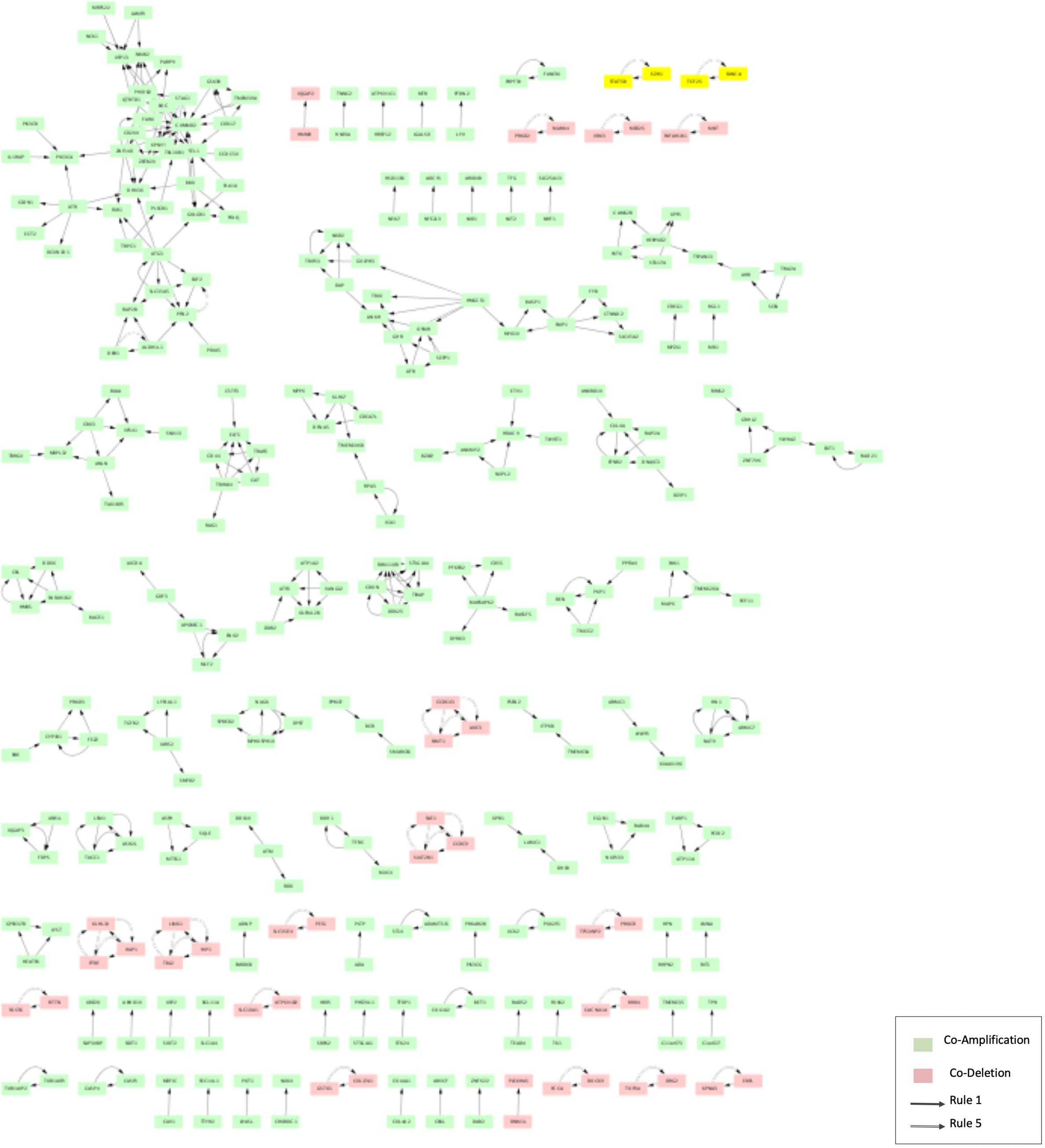


Figure 4.6 Representation of the Common Rules in the MSigDB Apoptosis Pathways with Rule Four Removed

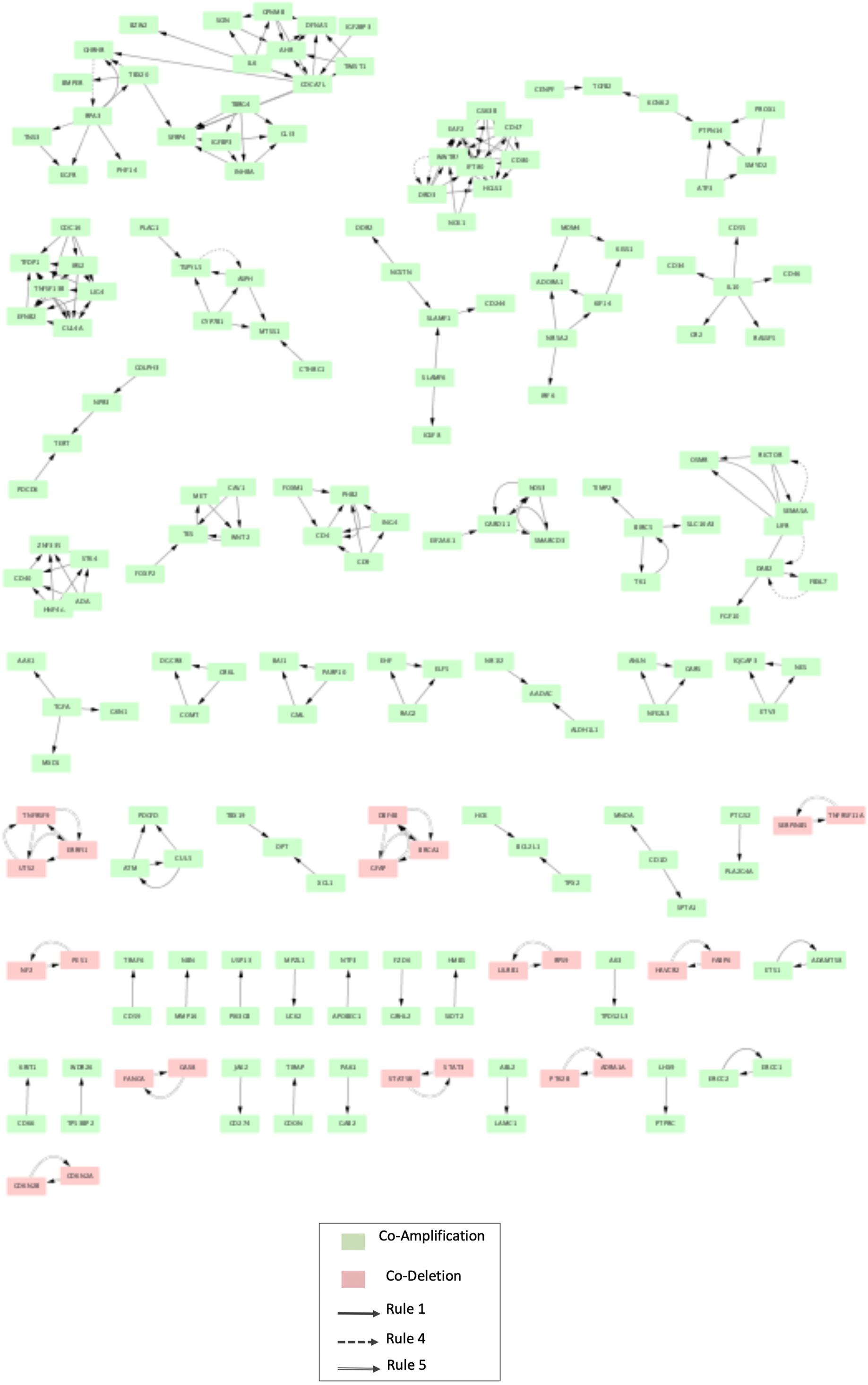


Figure 4.7 Representation of the Common Rules in the MSigDB Proliferation Pathways

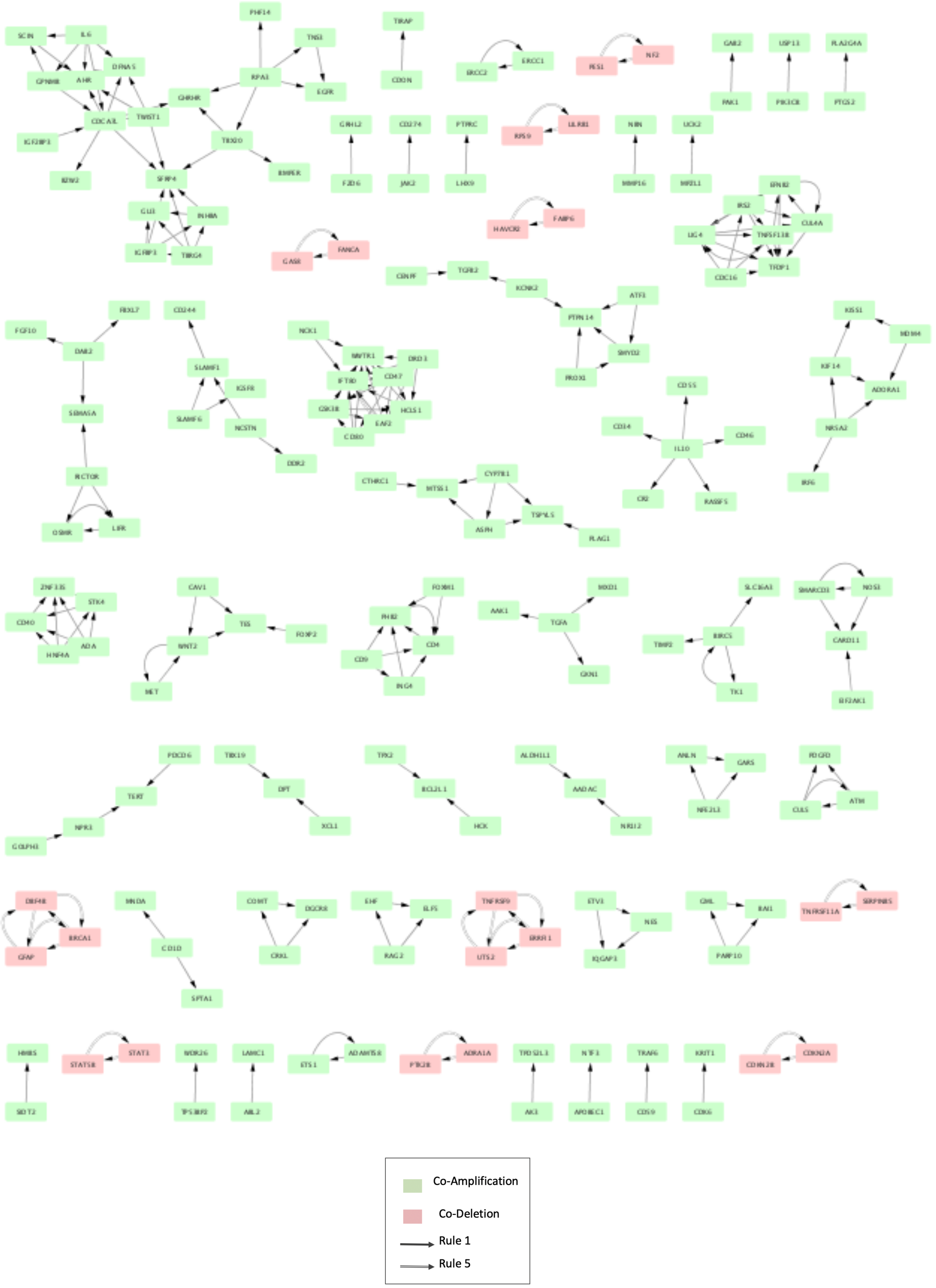


Figure 4.8 Representation of the Common Rules in the MSigDB Proliferation Pathways with Rule Four Removed

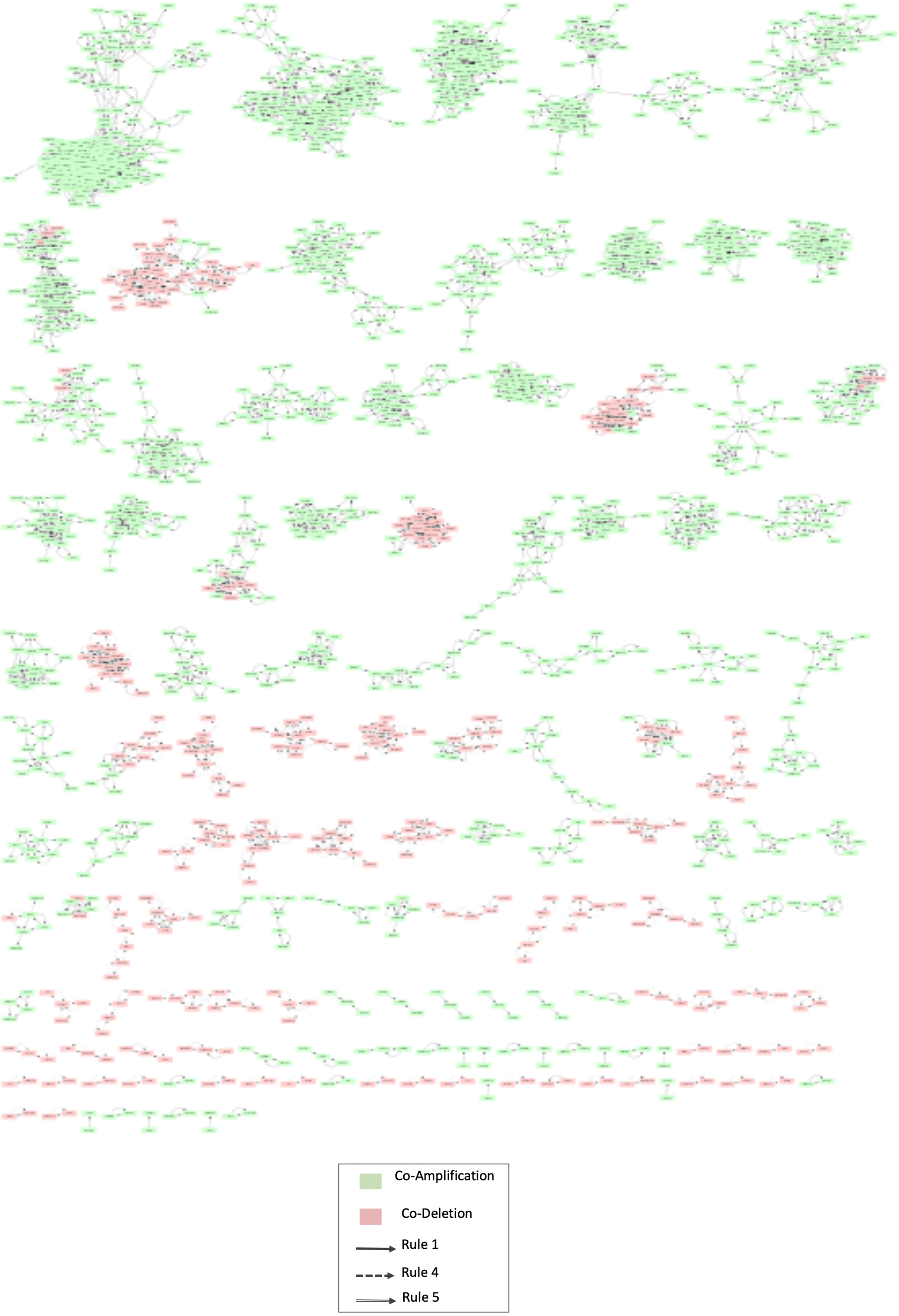


Figure 4.9 Representation of the Common Rules in the MSigDB Immunology Pathways

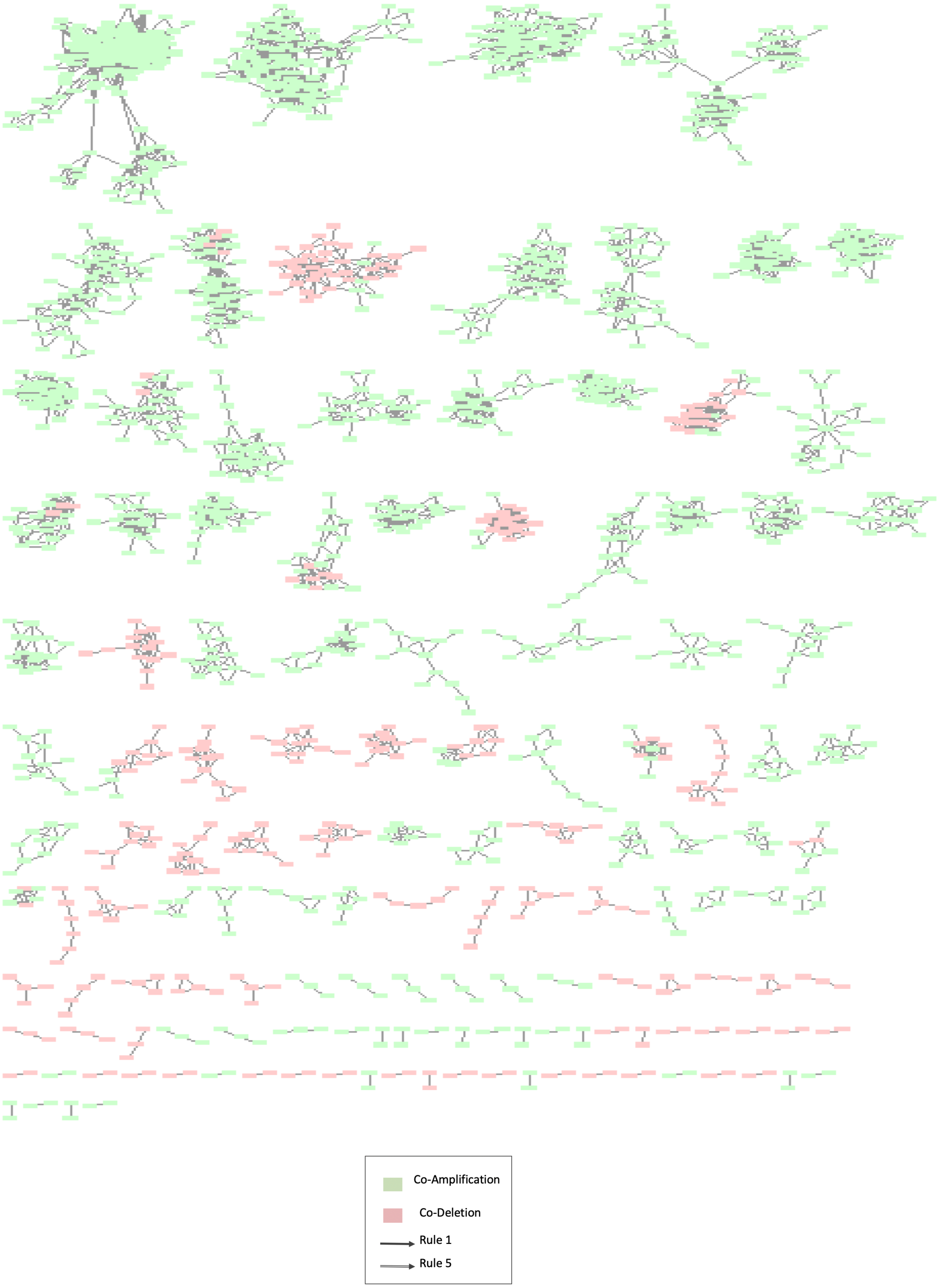


Figure 4.10 Representation of the Common Rules in the MSigDB Immunology Pathways with Rule Four Removed

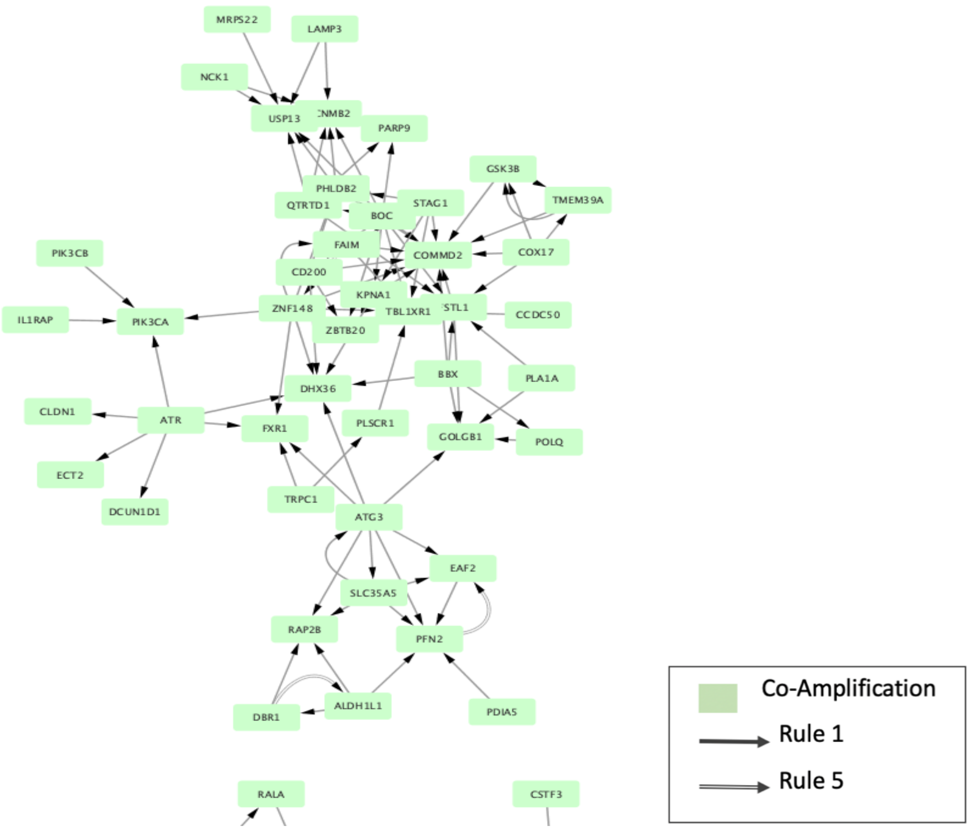


Figure 4.11 Magnification of a Sub-Network from Apoptosis Pathway Network

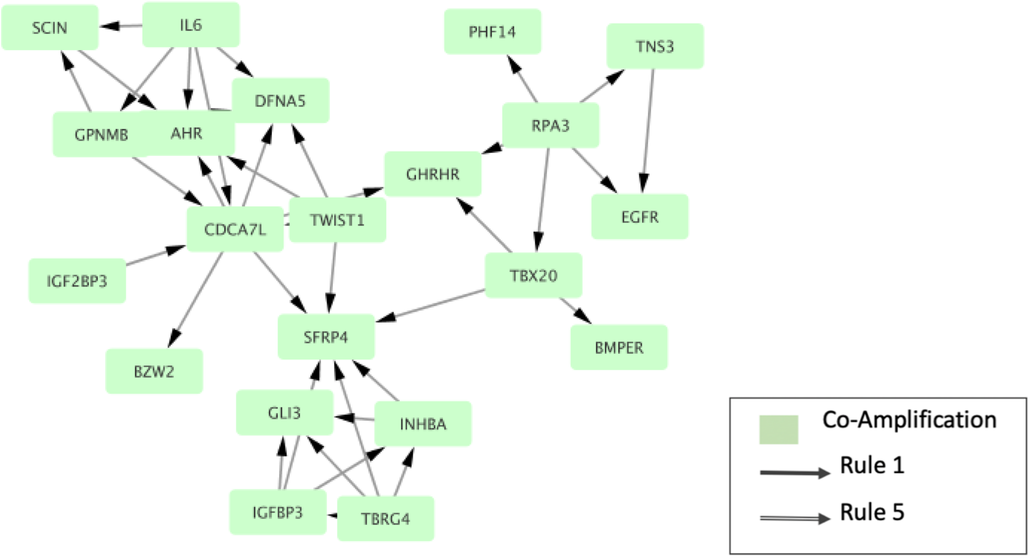


Figure 4.12 Magnification of a Sub-Network from Proliferation Pathway Network

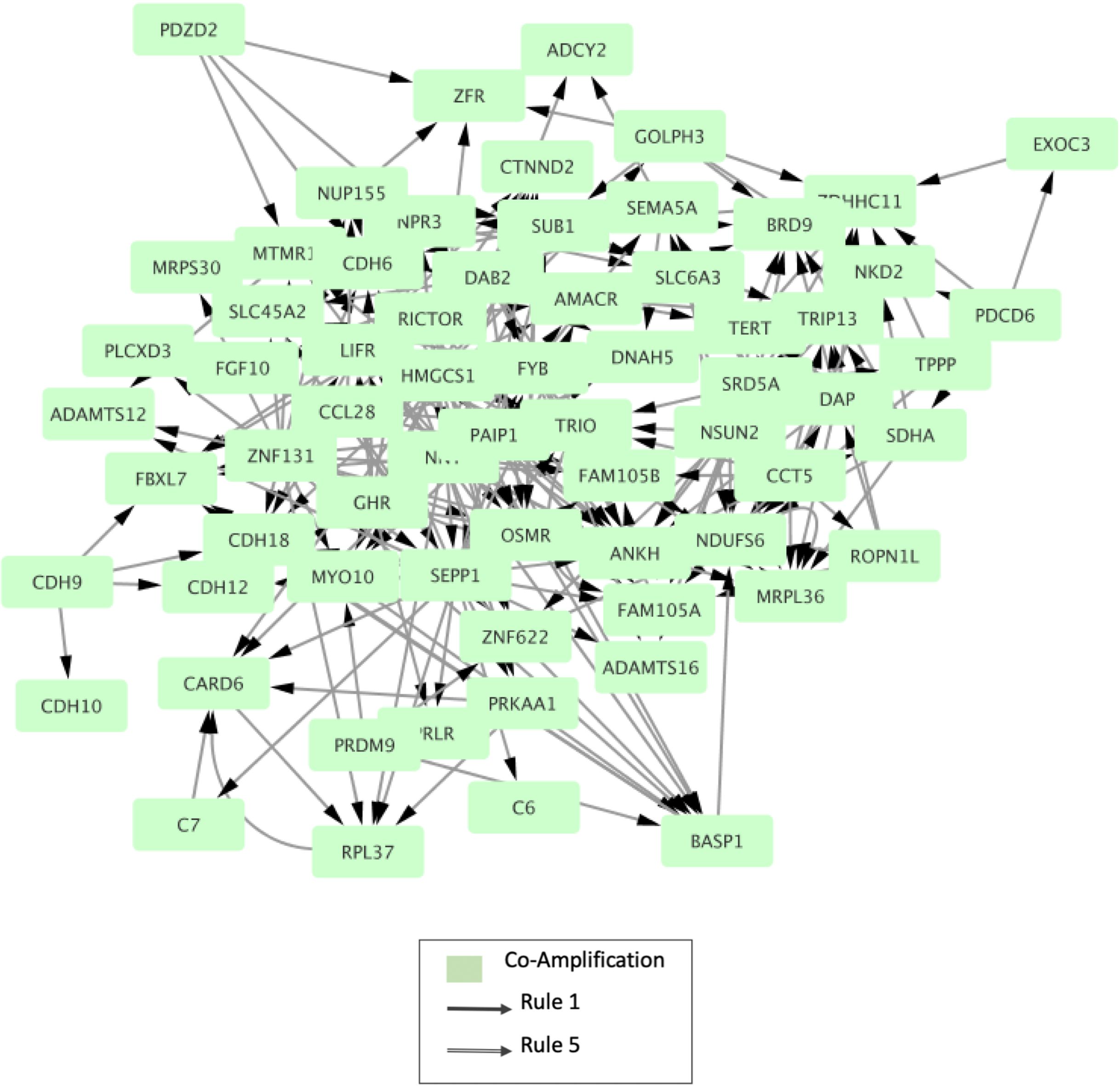


Figure 4.13 Magnification of a Sub-Network from Immunology Pathway Network

4.6 Analysis of PDL1 Interactions

PD1, present in the three patient cohorts as SNCA, and PDL1, present as CD274, have been shown to play a significant role in the immune system mechanisms of cancer cells, and have led to the development of immunotherapy drugs to treats many kinds of cancers. CTLA-4, similarly was also discovered to have a significant role in cell signaling and immune system mechanisms in the case of cancer cells. These two discoveries were recognized with the Nobel Prize in Medicine in 2018 (Boutros et al., 2016; Herbst et al., 2014). The interactions of PD1, PDL1 and CTLA-4 with other genes are of great biological interest, due to the role of these three genes in cancer mechanisms. We looked for Genet-CNV rules involving PDL1, PD1 and CTLA-4 that were common to two or more patient cohorts. We didn't find any interactions involving PD1 and CTLA-4 that were present in more than one patient cohort. The findings for PDL1 are summarized herein. The table below lists the rules involving the interaction of PDL1 found to be present in all three patient cohorts and validated with MSigDB:

Table 4.21 Interaction Rules for CD274 [PDL1]

Rule Type	Gene A	State	Gene B	State
Co-amplification	CD274	amplified	KIAA1432	amplified
Co-amplification	CD274	amplified	KIAA2026	amplified
Co-amplification	AK3	amplified	CD274	amplified
Co-amplification	INSL4	amplified	CD274	amplified
Co-amplification	JAK2	amplified	CD274	amplified
Co-amplification	KIAA1432	amplified	CD274	amplified

Table 4.22 CD274 [PDL1] Rules and their MSigDB Gene Sets

Rules		MSigDB
CD274	KIAA1432	C1, C2
CD274	KIAA2026	C1, C7
AK3	CD274	C1, C2, C7
INSL4	CD274	C1, C7
JAK2	CD274	C1, C2, C5, C7
KIAA1432	CD274	C1, C2

We provide the contingency tables for each of these gene pairs for each patient cohort. In these contingency tables, the number of patient samples for each of the four possible scenarios, A∧B, A∧¬B, ¬A∧B and ¬A∧¬B is provided.

Table 4.23 Contingency Table for PDL1 Rules from Patient Cohort GSE72194

GSE72194		Co-amplification Rules		A is amplified, B is amplified	A is amplified, B is not amplified	A is not amplified, B is amplified	A is not amplified, B is not amplified
Rule Type	**Error for Rule Type**	**Gene A**	**Gene B**	**A∧B**	**A∧¬B**	**¬A∧B**	**¬A∧¬B**
Co-amplification A⇒B	A∧¬B	CD274	KIAA1432	5	0	0	59
Co-amplification A⇒B	A∧¬B	CD274	KIAA2026	5	0	0	59
Co-amplification A⇒B	A∧¬B	AK3	CD274	3	0	2	59
Co-amplification A⇒B	A∧¬B	INSL4	CD274	5	0	0	59
Co-amplification A⇒B	A∧¬B	JAK2	CD274	3	0	2	59
Co-amplification A⇒B	A∧¬B	KIAA1432	CD274	5	0	0	59

Table 4.24 Contingency Table for PDL1 Rules from Patient Cohort GSE31800

GSE31800		Co-amplification Rules		A is amplified, B is amplified	A is amplified, B is not amplified	A is not amplified, B is amplified	A is not amplified, B is not amplified
Rule Type	**Error for Rule Type**	**Gene A**	**Gene B**	**A∧B**	**A∧¬B**	**¬A∧B**	**¬A∧¬B**
Co-amplification A⇒B	A∧¬B	CD274	KIAA1432	8	0	0	263
Co-amplification A⇒B	A∧¬B	CD274	KIAA2026	8	0	0	263
Co-amplification A⇒B	A∧¬B	AK3	CD274	8	0	0	263
Co-amplification A⇒B	A∧¬B	INSL4	CD274	8	0	0	263
Co-amplification A⇒B	A∧¬B	JAK2	CD274	8	0	0	263
Co-amplification A⇒B	A∧¬B	KIAA1432	CD274	8	0	0	263

Table 4.25 Contingency Table for PDL1 Rules from Patient Cohort GSE31800

GSE28572		Co-amplification Rules		A is amplified, B is amplified	A is amplified, B is not amplified	A is not amplified, B is amplified	A is not amplified, B is not amplified
Rule Type	**Error for Rule Type**	**Gene A**	**Gene B**	**A∧B**	**A∧¬B**	**¬A∧B**	**¬A∧¬B**
Co-amplification A⇒B	A∧¬B	CD274	KIAA1432	7	1	1	92
Co-amplification A⇒B	A∧¬B	CD274	KIAA2026	7	1	2	90
Co-amplification A⇒B	A∧¬B	AK3	CD274	5	1	3	91
Co-amplification A⇒B	A∧¬B	INSL4	CD274	4	1	4	91
Co-amplification A⇒B	A∧¬B	JAK2	CD274	5	1	3	90
Co-amplification A⇒B	A∧¬B	KIAA1432	CD274	7	1	1	91

PDL1 interactions common in any two datasets were also identified and visualized using Cytoscape.

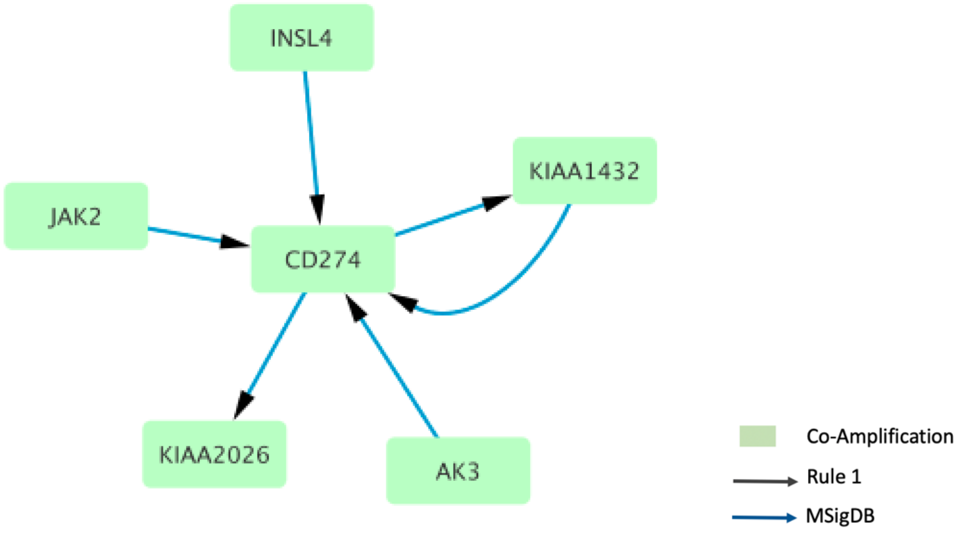


Fig 4.14 CD274 Interactions Common in all Datasets and Verified with MSigDB

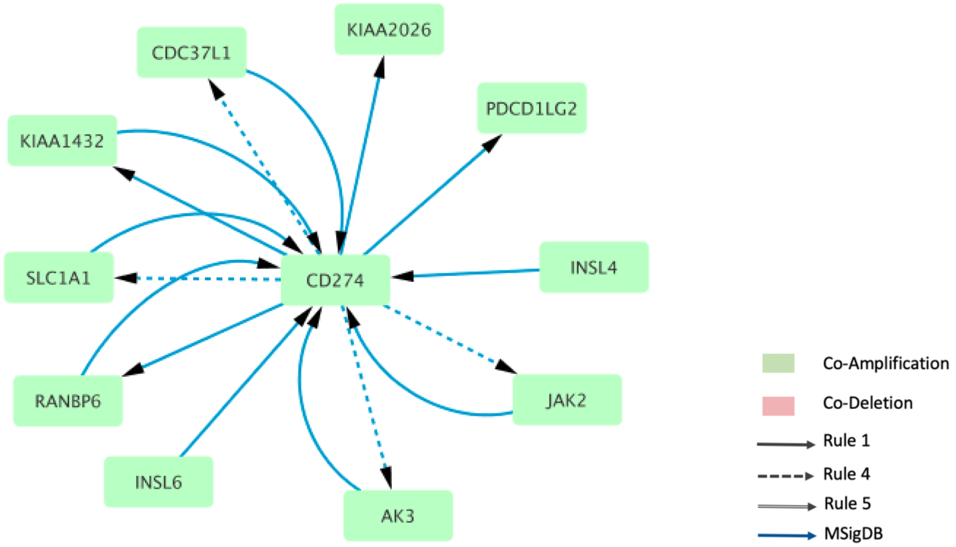


Fig 4.15 CD274 Interactions Common in GSE28572 and GSE72194 and Verified with MSigDB

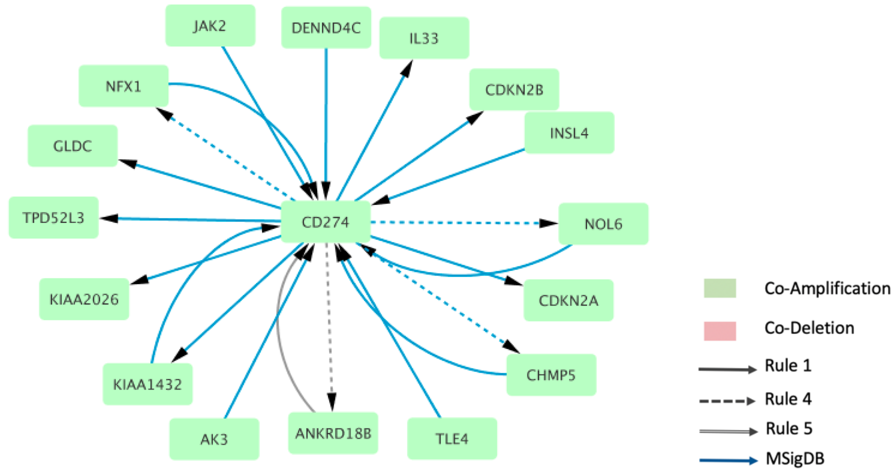


Fig 4.16 CD274 Interactions Common in GSE28572 and GSE31800 and Verified with MSigDB

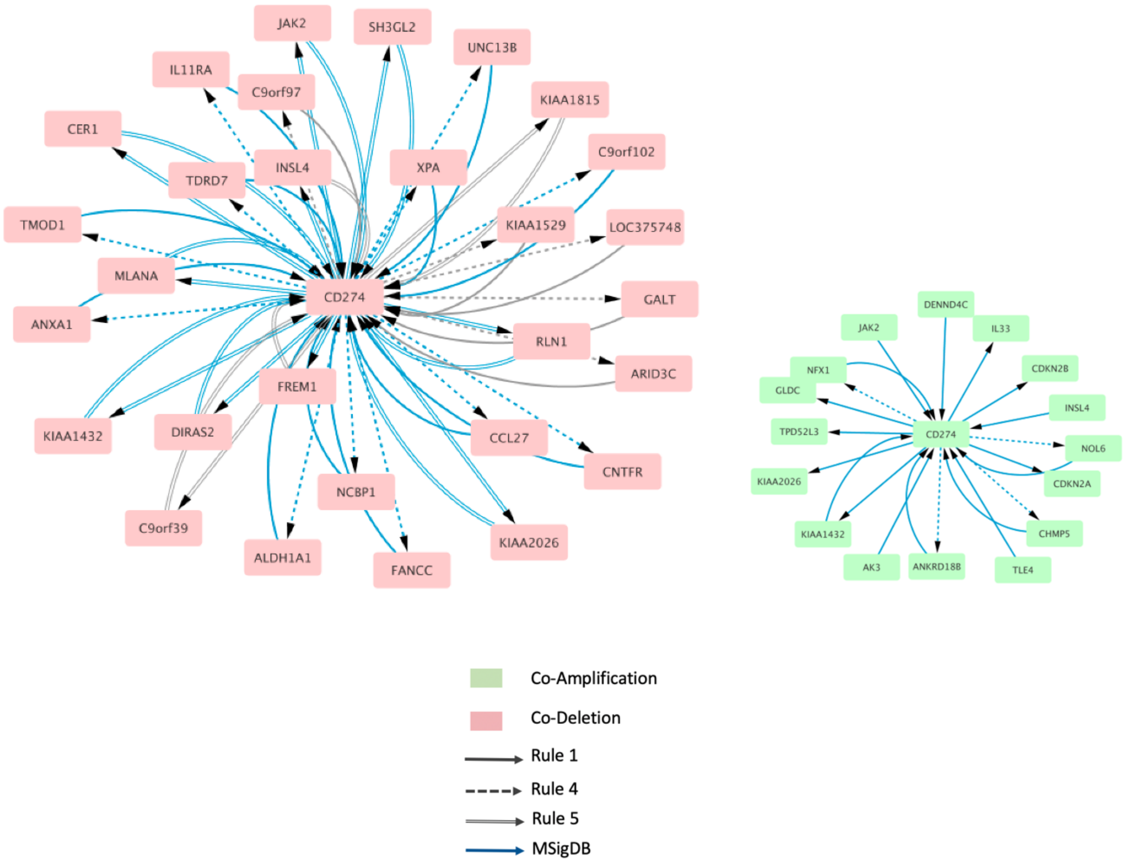


Fig 4.17 CD274 Interactions Common in GSE72194 and GSE31800 and Verified with MSigDB

4.7 Analysis of ZNF71 and CD27 Interactions

Out of the seven genes that were identified as prognostic biomarkers, CD27 and ZNF71 are of special significance. Therefore, the networks of ZNF71 and CD27 were analyzed. ZNF71 and CD27 were not present in all three datasets, but the rules present in any two datasets have been included in the networks of both genes. The figures also mark the rules in the networks validated with MSigDB.

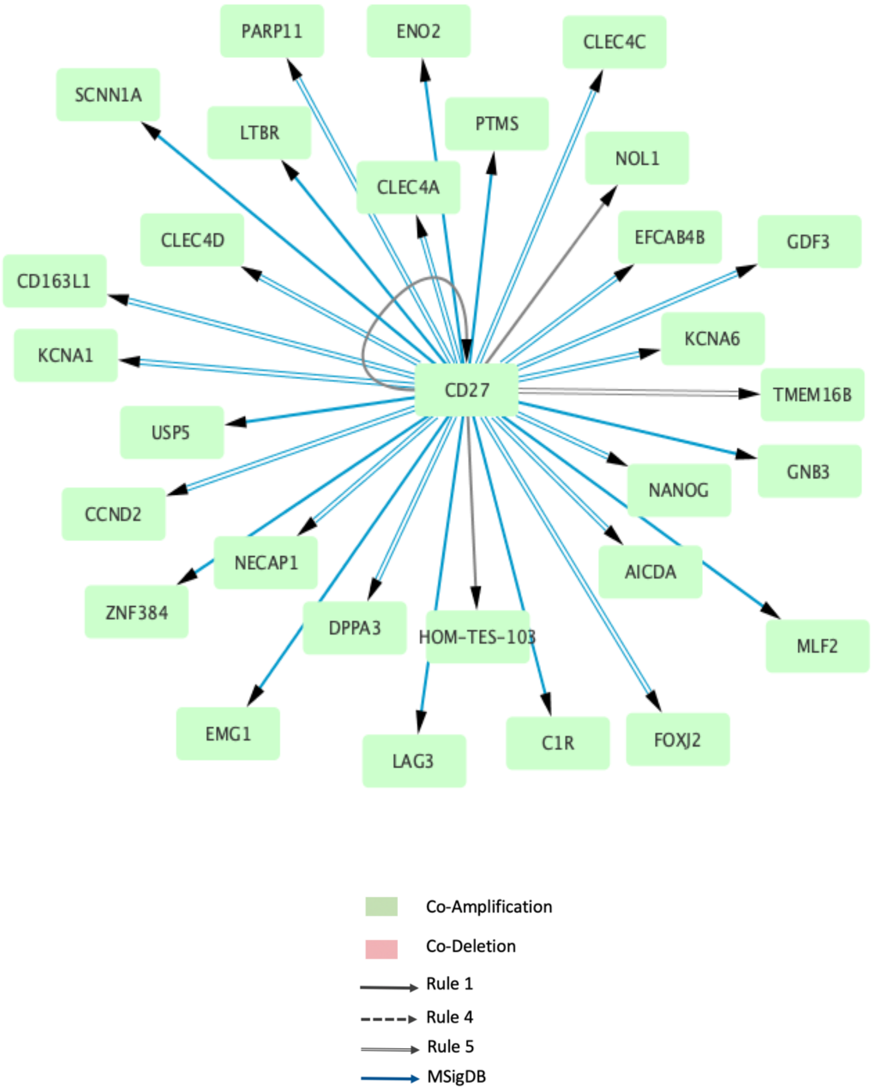


Figure 4.18 CD27 Interaction Network

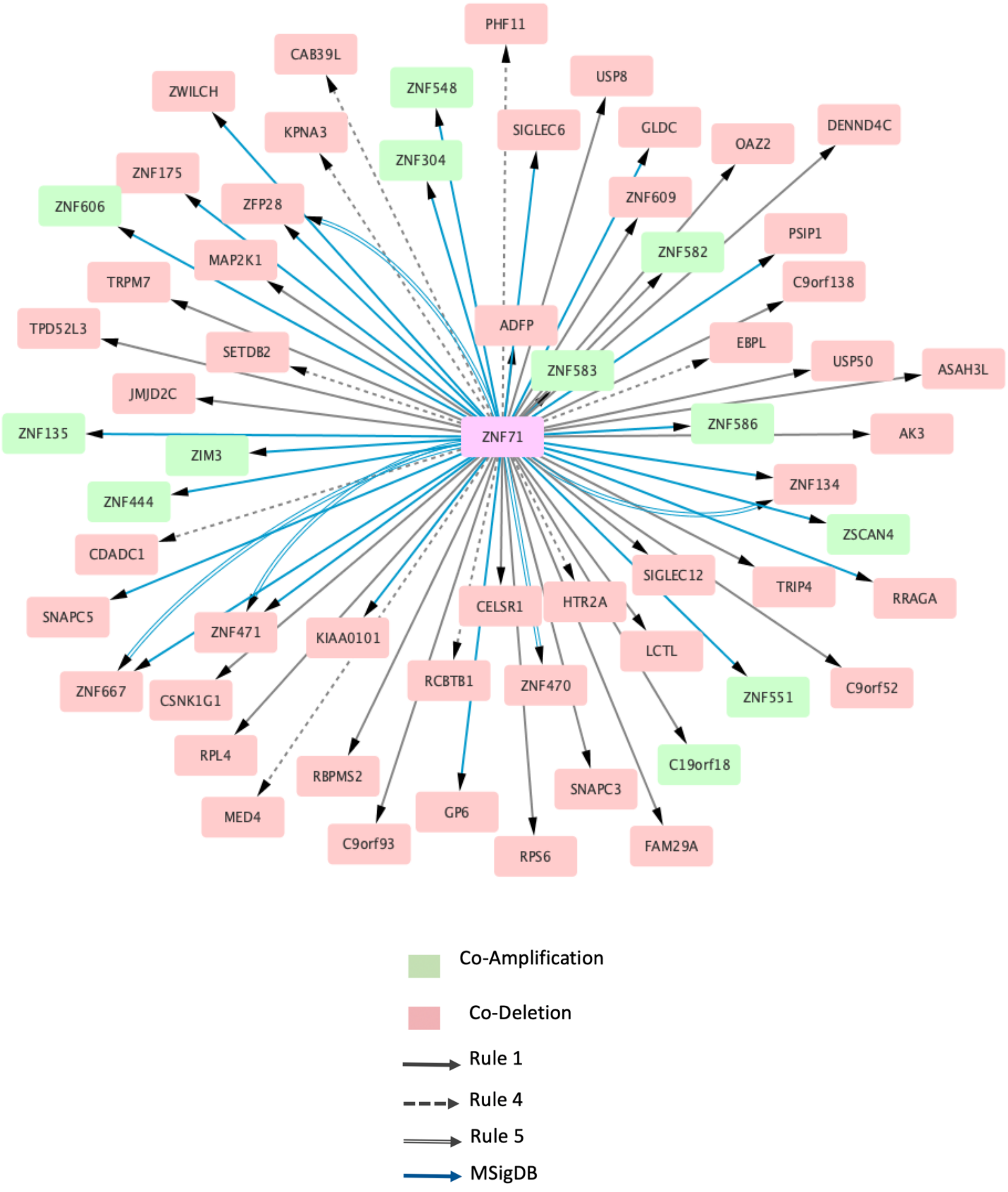


Figure 4.19 ZNF71 Interaction Network

4.10 Sensitivity, Specificity and Accuracy

The number of rules identified for seven genes were validated with MSigDB. The number of true positives, true negatives, false positives and false negatives was counted, and the accuracy was calculated for each of the three datasets.

In order to get the values for true positives, false positives and so on, we first calculated the total number of rules that could be obtained from the dataset. This was given by multiplying the total number of genes in a patient cohort, with the seven genes, and subtracting the seven possible rules where one of the seven genes was interacting with itself. True Positive refers to the unique rules identified by Genet-CNV that were also present in MSigDB. False Positive rules were those that were identified by Genet-CNV, but they were not present in MSigDB resources. False Negative rules were those that were not identified by Genet-CNV but were present in MSigDB. True Negative rules were those that were not identified by Genet-CNV as a rule and were also not present in MSigDB.

The formulae for calculating sensitivity, specificity and accuracy are given as follows:

$$\text{Sensitivity} = \frac{TP}{TP+FN}$$

$$\text{Specificity} = \frac{TN}{TN+FP}$$

$$\text{Accuracy} = \frac{TP+TN}{TP+TN+FP+FN}$$

where TP = True Positive, FP = False Positive, TN = True Negative and FN = False Negative.

The following tables shows the results obtained.

Table 4.26: Total Rules, Genet-CNV Rules and MSigDB Rules for Seven Genes

Dataset	Total Possible Rules	Unique Rules from Genet-CNV	Rules in MSigDB (TP)
GSE31800	138033	11176	6283
GSE72194	96762	22981	11703
GSE28572	50037	2954	1055

Table 4.27: Sensitivity, Specificity and Accuracy Levels for Seven Genes

Dataset	True Positive	False Positive	True Negative	False Negative	Sensitivity	Specificity	Accuracy
GSE31800	6283	4893	56505	70352	0.08	0.92	0.45
GSE72194	11703	11278	35829	37952	0.24	0.76	0.49
GSE28572	1055	1899	27352	19731	0.05	0.94	0.57

We also checked our calculations to see if on adding up all the true positives, true negatives, false positives and false negatives, we get the total possible number of rules. We found that the sum of TP, TN, FP and FN rules was always equal to the total number of rules possible, does validating our calculations.

Table 4.28: Comparison of All Possible Rules with Rules Counted in All Four Categories

Dataset	Total Possible Rules	Sum of TP, FP, TN and FN
GSE31800	138033	138033
GSE72194	96762	96762
GSE28572	50037	50037

4.9 Comparison with BooleanNet

Using Genet-CNV, the accuracy is around 50 per cent for each patient cohort. This is significantly better than the results obtained using BooleanNet, which does not evaluate a single rule for any of the three patient cohorts.

BooleanNet is the algorithm developed by Sahoo et al (Sahoo et al., 2008) referenced in section 2.2.1 This algorithm uses a variation of the Boolean Implication network algorithm to count the number of Boolean implication rules found in genome wide gene expression data from microarrays. The details of the implementation of the algorithm are mentioned in section 2.2.1.

We did a run of BooleanNet on our data input files and did not find any rules for any of the three datasets. We also used a sample file available with BooleanNet to test both Genet-CNV and BooleanNet (Gene Expression Omnibus).

Table 4.29: Comparison of Rules from BooleanNet and Genet-CNV

Dataset	No. of Rules with BooleanNet/Rules Confirmed with MSigDB	Accuracy	No. of Unique Rules with Genet CNV/Rules Confirmed with MSigDB	Accuracy with Genet-CNV
GSE31800	0/0	0	11176/6283	0.45
GSE72194	0/0	0	22981/11703	0.49
GSE28572	0/0	0	2954/1055	0.57
Sample File given with BooleanNet	1841/-	-	47/-	-

DNA CNV is an anomaly, and as such, only a very small percentage of the samples analyzed will have an amplified or deleted state for any given gene. Most of the samples will have a normal state, denoted by 0. With gene expression datasets, however, most samples will show a high or low value, denoting upregulation or downregulation respectively for any given genes, and very few samples will have an intermediate value (0) for any given gene.

As a result, the dataset we are using, representing DNA CNV data, make up very sparse matrices, with most values as 0, and very few as 1 (amplification) or -1 (deletion). The input data used in BooleanNet, being gene expression data, will have a much denser matrix, with most values as 1 or -1. The inability of BooleanNet to yield any rules for the datasets used in this study can be cited to the sparseness of the input matrices. With the same dataset, Genet-CNV identifies thousands of rules, thus proving to be much more sensitive than BooleanNet for DNA-CNV data.

It took BooleanNet 19 minutes to go through GSE31800, whereas it took Genet-CNV 12 minutes to go through the same dataset on a macOS Mojave with a 2.3GHz Intel Core i5 processor and 8 GB 2133 MHz LPDDR3 RAM. Therefore, Genet-CNV is also faster than BooleanNet.

Chapter 5

Conclusions and Future Work

Genet-CNV is a computational framework for modelling genome wide co-occurrences of DNA CNV data, and as a development over its previous implementation, it has eliminated the constraint that allows only binomial variable to be used. Multinomial variables can be represented as the variable and its complement, and therefore can be generalized to further analyses.

NSCLC is a very complicated disease to establish causal relations for, and the need for molecular network analysis tools that can indicate cellular pathways and gene interactions in various cellular processes, which might be of interest to biologists and clinicians. Genet-CNV is an effective framework to carry out such analyses. Furthermore, the results obtained by Genet-CNV across three patient cohorts have been mined and analyzed for common associations, and the common associations can serve as launching points for biological investigation. The validation of rules with the benchmark dataset, MSigDB, further establishes the significance of an association, and the results thus generated can yield various insights for investigating disease mechanisms and therapeutic targets.

In the future, we intend to compare the performance of Genet-CNV with that of a modified implementation of BooleanNet, which also incorporates the usage of Fisher's Test (Sinha 2014). Currently, the association networks of ZNF71, an important prognosis biomarker obtained by Genet-CNV are being used to generate hypotheses and seek biological validation from them in Guo Labs at the WVU Cancer Institute. Another direction of future work is to develop an algorithm to carry out integrated analysis of genome wide DNA CNV and gene expression data to discover genes of importance for clinical and therapeutic processes. Another student in Dr. Guo's laboratory is seeking to develop this integrated analysis computational framework.

REFERENCES

Aramburu, A., Zudaire, I., Pajares, M. J., Agorreta, J., Orta, A., Lozano, M. D., . . . Montuenga, L. M. (2015). Combined clinical and genomic signatures for the prognosis of early stage non-small cell lung cancer based on gene copy number alterations. *BMC Genomics, 16*, 752. doi:10.1186/s12864-015-1935-0

Blandin Knight, S., Crosbie, P. A., Balata, H., Chudziak, J., Hussell, T., & Dive, C. (2017). Progress and prospects of early detection in lung cancer. *Open biology, 7*(9), 170070. doi:10.1098/rsob.170070

Çakır, M., Binder, H, Wirth, H. (2016). Profiling of Genetic Switches using Boolean Implications in Expression Data. *Journal of Integrative Bioinformatics*. doi:https://doi.org/10.1515/jib-2014-246

Calvano, S. E., Xiao, W., Richards, D. R., Felciano, R. M., Baker, H. V., Cho, R. J., . . . Host Response to Injury Large Scale Collab. Res, P. (2005). A network-based analysis of systemic inflammation in humans. *Nature, 437*(7061), 1032-1037. doi:10.1038/nature03985

Carter, N. P. (2007). Methods and strategies for analyzing copy number variation using DNA microarrays. *Nat Genet, 39*(7 Suppl), S16-21. doi:10.1038/ng2028

Chuang, H. Y., Lee, E., Liu, Y. T., Lee, D., & Ideker, T. (2007). Network-based classification of breast cancer metastasis. *Mol Syst Biol, 3*, 140. doi:10.1038/msb4100180

Dymacek, J. M., Snyder-Talkington, B. N., Raese, R., Dong, C., Singh, S., Porter, D. W., . . . Guo, N. L. (2018). Similar and Differential Canonical Pathways and Biological Processes Associated With Multiwalled Carbon Nanotube and Asbestos-Induced Pulmonary Fibrosis: A 1-Year Postexposure Study. *Int J Toxicol, 37*(4), 276-284. doi:10.1177/1091581818779038

Emilsson, V., Thorleifsson, G., Zhang, B., Leonardson, A. S., Zink, F., Zhu, J., . . . Stefansson, K. (2008). Genetics of gene expression and its effect on disease. *Nature, 452*(7186), 423-428. doi:10.1038/nature06758

Genetics Home Reference. (2019, February 3). What are single nucleotide polymorphisms (SNPs)? Retrieved from https://ghr.nlm.nih.gov/primer/genomicresearch/snp

Goldstraw, P., Chansky, K., Crowley, J., Rami-Porta, R., Asamura, H., Eberhardt, W. E., . . . Participating, I. (2016). The IASLC Lung Cancer Staging Project: Proposals for Revision of the TNM Stage Groupings in the Forthcoming (Eighth) Edition of the TNM Classification for Lung Cancer. *J Thorac Oncol, 11*(1), 39-51. doi:10.1016/j.jtho.2015.09.009

Guo, L., Cukic, B., & Singh, H. (2003). Predicting Fault Prone Modules by the Dempster-Shafer Belief Networks. *Proc IEEE Int Autom Softw Eng Conf, 2003*, 249-252. doi:10.1109/ASE.2003.1240314

Guo, N. L., Dowlati, A., Raese, R. A., Dong, C., Chen, G., Beer, D. G., . . . Qian, Y. (2018). A Predictive 7-Gene Assay and Prognostic Protein Biomarkers for Non-small Cell Lung Cancer. *EBioMedicine, 32*, 102-110. doi:10.1016/j.ebiom.2018.05.025

Guo, N. L., & Wan, Y. W. (2012). Pathway-based identification of a smoking associated 6-gene signature predictive of lung cancer risk and survival. *Artif Intell Med, 55*(2), 97-105. doi:10.1016/j.artmed.2012.01.001

Herbst, R. S., Heymach, J. V., & Lippman, S. M. (2008). Lung cancer. *N Engl J Med, 359*(13), 1367-1380. doi:10.1056/NEJMra0802714

Jansen, R., Yu, H., Greenbaum, D., Kluger, Y., Krogan, N. J., Chung, S., . . . Gerstein, M. (2003). A Bayesian networks approach for predicting protein-protein interactions from genomic data. *Science, 302*(5644), 449-453. doi:10.1126/science.1087361

Khojasteh, M., Lam, W. L., Ward, R. K., & MacAulay, C. (2005). A stepwise framework for the normalization of array CGH data. *BMC Bioinformatics, 6*, 274. doi:10.1186/1471-2105-6-274

Liberzon, A., Birger, C., Thorvaldsdottir, H., Ghandi, M., Mesirov, J. P., & Tamayo, P. (2015). The Molecular Signatures Database (MSigDB) hallmark gene set collection. *Cell Syst, 1*(6), 417-425. doi:10.1016/j.cels.2015.12.004

Liberzon, A., Subramanian, A., Pinchback, R., Thorvaldsdottir, H., Tamayo, P., & Mesirov, J. P. (2011). Molecular signatures database (MSigDB) 3.0. *Bioinformatics, 27*(12), 1739-1740. doi:10.1093/bioinformatics/btr260

Micke, P., Edlund, K., Holmberg, L., Kultima, H. G., Mansouri, L., Ekman, S., . . . Botling, J. (2011). Gene copy number aberrations are associated with survival in histologic subgroups of non-small cell lung cancer. *J Thorac Oncol, 6*(11), 1833-1840. doi:10.1097/JTO.0b013e3182295917

Milo, R., Shen-Orr, S., Itzkovitz, S., Kashtan, N., Chklovskii, D., & Alon, U. (2002). Network motifs: simple building blocks of complex networks. *Science, 298*(5594), 824-827. doi:10.1126/science.298.5594.824

Mitchell, T. M. (1997). *Machine Learning*. New York: McGraw-Hill.

Nesbitt, J. C., Putnam, J. B., Jr., Walsh, G. L., Roth, J. A., & Mountain, C. F. (1995). Survival in early-stage non-small cell lung cancer. *Ann Thorac Surg, 60*(2), 466-472.

Sachs, K., Perez, O., Pe'er, D., Lauffenburger, D. A., & Nolan, G. P. (2005). Causal protein-signaling networks derived from multiparameter single-cell data. *Science, 308*(5721), 523-529. doi:10.1126/science.1105809

Sahoo, D., Dill, D. L., Gentles, A. J., Tibshirani, R., & Plevritis, S. K. (2008). Boolean implication networks derived from large scale, whole genome microarray datasets. *Genome Biol, 9*(10), R157. doi:10.1186/gb-2008-9-10-r157

Sahoo, D., Dill, D. L., Tibshirani, R., & Plevritis, S. K. (2007). Extracting binary signals from microarray time-course data. *Nucleic Acids Res, 35*(11), 3705-3712. doi:10.1093/nar/gkm284

Scientific, T. F. (2019, January 13). Analysis Power Tools (APT) -- Release 2.10.2.2. Retrieved from http://media.affymetrix.com/support/developer/powertools/changelog/index.html

Shah, R., Sabanathan, S., Richardson, J., Mearns, A. J., & Goulden, C. (1996). Results of surgical treatment of stage I and II lung cancer. *J Cardiovasc Surg (Torino), 37*(2), 169-172.

Sinha, S., Tsang, EK, Zeng, H, Meister, M, Dill, DL (2014). Mining TCGA Data Using Boolean Implications. *PLOS ONE*. doi:https://doi.org/10.1371/journal.pone.0102119

Solinas-Toldo, S., Lampel, S., Stilgenbauer, S., Nickolenko, J., Benner, A., Dohner, H., . . . Lichter, P. (1997). Matrix-based comparative genomic hybridization: biochips to screen for genomic imbalances. *Genes Chromosomes Cancer, 20*(4), 399-407.

Starczynowski, D. T., Lockwood, W. W., Delehouzee, S., Chari, R., Wegrzyn, J., Fuller, M., . . . Karsan, A. (2011). TRAF6 is an amplified oncogene bridging the RAS and NF-kappaB pathways in human lung cancer. *J Clin Invest, 121*(10), 4095-4105. doi:10.1172/JCI58818

Torre, L. A., Bray, F., Siegel, R. L., Ferlay, J., Lortet-Tieulent, J., & Jemal, A. (2015). Global cancer statistics, 2012. *CA Cancer J Clin, 65*(2), 87-108. doi:10.3322/caac.21262

van de Wiel, M. A., Kim, K. I., Vosse, S. J., van Wieringen, W. N., Wilting, S. M., & Ylstra, B. (2007). CGHcall: calling aberrations for array CGH tumor profiles. *Bioinformatics, 23*(7), 892-894. doi:10.1093/bioinformatics/btm030

van Wieringen, W. N., Belien, J. A., Vosse, S. J., Achame, E. M., & Ylstra, B. (2006). ACE-it: a tool for genome-wide integration of gene dosage and RNA expression data. *Bioinformatics, 22*(15), 1919-1920. doi:10.1093/bioinformatics/btl269

van Wieringen, W. N., Unger, K., Leday, G. G., Krijgsman, O., de Menezes, R. X., Ylstra, B., & van de Wiel, M. A. (2012). Matching of array CGH and gene expression microarray features for the purpose of integrative genomic analyses. *BMC Bioinformatics, 13*, 80. doi:10.1186/1471-2105-13-80

Venkatraman, E. S., & Olshen, A. B. (2007). A faster circular binary segmentation algorithm for the analysis of array CGH data. *Bioinformatics, 23*(6), 657-663. doi:10.1093/bioinformatics/btl646

Walters, S., Maringe, C., Coleman, M. P., Peake, M. D., Butler, J., Young, N., . . . Group, I. M. W. (2013). Lung cancer survival and stage at diagnosis in Australia, Canada, Denmark, Norway, Sweden and the UK: a population-based study, 2004-2007. *Thorax, 68*(6), 551-564. doi:10.1136/thoraxjnl-2012-202297

Wang, K., Li, M., Hadley, D., Liu, R., Glessner, J., Grant, S. F., . . . Bucan, M. (2007). PennCNV: an integrated hidden Markov model designed for high-resolution copy number variation detection in whole-genome SNP genotyping data. *Genome Res, 17*(11), 1665-1674. doi:10.1101/gr.6861907

Boutros, C., Tarhini, A., Routier, E., Lambotte, O., Ladurie, F. L., Carbonnel, F., . . . Robert, C. (2016). Safety profiles of anti-CTLA-4 and anti-PD-1 antibodies alone and in combination. Nat Rev Clin Oncol, 13(8), 473-486. doi:10.1038/nrclinonc.2016.58

Gene Expression Omnibus. Gene Expression Omnibus. Retrieved from https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE119087

Herbst, R. S., Soria, J. C., Kowanetz, M., Fine, G. D., Hamid, O., Gordon, M. S., . . . Hodi, F. S. (2014). Predictive correlates of response to the anti-PD-L1 antibody MPDL3280A in cancer patients. Nature, 515(7528), 563-567. doi:10.1038/nature14011

PennCNV. (2019). Affymetrix CNV Calling oOerview. Retrieved from http://penncnv.openbioinformatics.org/en/latest/user-guide/affy/

Liu, L, Desmarais, MC, (1997) A method of learning implication networks from empirical data: algorithm and Monte-Carlo simulation-based validation. IEEE Transactions on Knowledge and Data Engineering, DOI**:** 10.1109/69.649321