NOTTINGHAM TRENT UNIVERSITY

In Silico Modelling of the TP53 Pathway in Cancers Using Artificial Neural Network based Systems Biology Approaches

Dalia Mehaisi

A thesis submitted in partial fulfilment of the requirements of Nottingham Trent University for the degree of Doctor of Philosophy (PhD)

March 2024

Copyright Statement

Dalia Mehaisi

(N0763646)

Abstract

Cancer, a major health issue and one of the most common causes of death worldwide, arises through a multi-stage process that involves several genetic alterations (pathological, immunological, and physiological). Researchers are continually seeking to explore such alterations at the molecular level to gain knowledge that can be used for disease management and prevention, resulting in several large-scale transcriptomic technologies to estimate whole genome expression profiles for cancer. However, such analytical approaches generate massive volumes of data, which need careful processing to extract meaningful information using statistical and computational approaches. Some of these approaches have been dedicated to studying cancer through interrogation of pathway models based on molecular data and based on mining of the literature corpus to obtain deep insights which could help in drug discovery and the achievement of personalized medicine for cancer. These methods tend to address the dimensionality and complexity issues associated with large-scale technologies by presenting the data using signalling network models and pathway knowledge graphs. However, the possibility of identifying novel interactions and disease drivers remains limited, as most of these approaches are based on knowledge obtained from the literature through manual curation.

ANN-based integrative data mining approaches have been successful in cancer research, coping with noise and dimensionality associated with high throughput data, allowing for the identification of novel interactions and drivers related to diseases. These drivers can be used as a panel for the classification of certain conditions or as targets for new therapeutic interventions.

This project applies ANN approaches for pathway data mining through a series of analyses leading to the identification of key interactions associated with the TP53 pathway in cancer. The first analysis indicates the novel drivers associated with the TP53 pathway in colorectal cancer. The second analysis suggests common and unique predictors associated with the TP53 pathway in the Mutant- and Wild-type status of the *TP53* gene using three cohorts: colon and rectum cancer (COADREAD), pancreatic cancer (PAAD), and stomach cancer (STAD) from cases in The Cancer Genome Atlas (TCGA).

This analysis also identified a panel of differential drivers associated with theTP53 pathway in the Missense mutation status of the TP53 gene for the investigated cohorts. The study integrates the findings and compares the ANN driver results with the existing pathway analysis tool, MetaCore. The final analysis revealed a panel of differential drivers associated with the TP53 pathway in the Wild-type state of the TP53 gene for the studied cohorts.

Key terms: **Transcriptomic data[1], Artificial Neural Network[2], pathway modelling[3], TP53 pathway[4], Predictors[5], Drivers[6], MetaCore[7].**

Key terms descriptions: 1- Collective information of RNA transcripts, 2- Computational biology technique used for data analysis, 3- A Guideline represent the order and the relationship of molecules involved in a certain cellular process, 4- A crucial signalling pathway involved in regulating cellular responses during stresses, 5- Inputs that used to indicate certain future outcome, 6- Inputs that lead or operate a certain cellular response in a specific situation, 7- A platform that assist scientist in analysis and visualization of genomic data.

Acknowledgment

In the name of Allah, the most Generous and the most Merciful, all praises to Allah, the Almighty, and Compassionate, for the numerous blessings, patience, health, and support to complete this academic work.

I offer my acknowledgment and my thanks to the director of the study, Professor Graham Ball, for giving me this fantastic opportunity and for his valuable guidance throughout the project. I would also like to thank the co-supervisors, Dr. David Boocock and Dr. Sergio Rutella, and the independent assessor Dr. Amanda Coutts for their encouragement, support, and critical insights. Many thanks to Dr. Dimitrios Zafeiris for sharing knowledge and expertise, especially in the technical work. I thank all of them for their thoughtful comments and suggestions.

Many thanks to the IT technical team for their cooperation and support in the difficulties associated with the technical side of the project.

My sincere gratitude goes to my family and my friends for their motivation and support in every aspect of my life.

Table of Contents

List of Figures

ORIGINAL LIST (TO BE DELETED)

List of Tables

List of Abbreviations

ANN	Artificial neural network
ANNI	Artificial neural network inference
ANOVA	Analysis of variance
APC	Adenomatous polyposis coli
BP	Backpropagation networks
CIN	Chromosomal instability
COADREAD	Colon and rectum cancer
CRC	Colorectal cancer
DNA	Deoxyribonucleic acid
FCS	Functional class scoring
GO	Gene ontology
GSEA	Gene set enrichment analysis
HNPCC	Hereditary non-polyposis colorectal cancer
IHC	Immunohistochemistry
IR	Ionizing radiation
KEGG	Kyoto Encyclopedia of Genes and Genomes
MCCV	Monte Carlo cross-validation
MLP	Multilayer perceptron
mRNA	Messenger ribonucleic acid

ODE	Ordinary differential equation model
ORA	Over-representation analysis
PAAD	Pancreatic cancer
PCA	Principal component analysis
PCR	Polymerase chain reaction
PPI	Protein-protein interaction network
PT	Pathway topology
PT	Pathway topology based methods
RMS	Root mean square value
RPKM	Reads per kilobase of exon model per million mapped reads
SOM	Self-organizing map
STAD	Stomach cancer
STRING	Search Tool for the Retrieval of Interacting Genes/Proteins
SVM	Support vector machine
TCGA	The Cancer Genome Atlas
UC	Ulcerative colitis

CHAPTER 1
GENERAL INTRODUCTION

1.1. Cancer

Cancer is a group of diseases arising from genetic alterations that occur in cells, causing dysregulation in key signalling pathways, with a consecutive formation of new cellular properties. Such appearances manifest excessive proliferation and resistance to cell death, which result in cancer initiation and progression (Hanahan & Weinberg, 2011). Most genetic alterations occur as mutations that lead to the activation of growth control genes (oncogenes) and loss of function in tumour-suppressor genes. Other modifications may involve DNA stability genes, which play an indirect role in tumourigenesis by increasing the mutations of other genes (Sever & Brugge, 2010). Specific oncogenes and tumour suppressor genes are continually being discovered and linked to various types of cancers. Emerging genetic information, and knowledge arising from it, enhance the understating of molecular mechanisms that lead to cancer, facilitating the introduction of new targeted therapies for cancer treatment. These targeted therapeutic protocols have enabled some tangible improvements and are of increasing importance in clinical practice, spurring increasing cancer research in the field of genome analysis (Yan et al., 2011).

High-throughput technologies, such as DNA microarray and RNA Sequencing, allow the analysis of cancer genes on a large scale, assisting in identifying relationships between genes and building pathway models to gain insights into the mechanism of disease formation and progression. However, the interpretation of genomic data needs careful consideration due to the complexity and non-linearity of the techniques (Bernard & Wittwer, 2002). Several computational databases and statistical analysis tools have been developed to facilitate the interpretation of genetic information to improve clinical practice for cancer patients (Zou et al., 2015).

1.2. Gene expression analysis

Gene expression analysis is the process involving the measurement of the expressed genes within the cell at a specific time. The expression of the genes requires several regulations, most of which occur during the transcriptional level; hence the expression of genes is used to indicate protein functions, with significant applications in cancer identification and classification. Researchers can distinguish cancer cells from normal cells based on the differences in the expression level of specific genes. They can also discover genetic signatures, which can help in diagnosis, prognosis, and therapy prediction (Russo et al., 2003).

Different techniques can be used for gene expression profiling, which measures the amount of mRNA in the cells and uses it to indicate the transcribed genes. Figure 1.1 Provides a general overview of the most common methods. One of the formal methods is the polymerase chain reaction (PCR), which has been used to amplify a specific gene of interest and link it to a particular type of cancer. The method was then developed to reverse transcriptase (RTPCR) and then to real-time quantitative reverse transcriptase (real-time qRT-PCR). Although these methods are reliable and easy to apply in different research settings, they have a limited view of the cancer genomic complete picture.

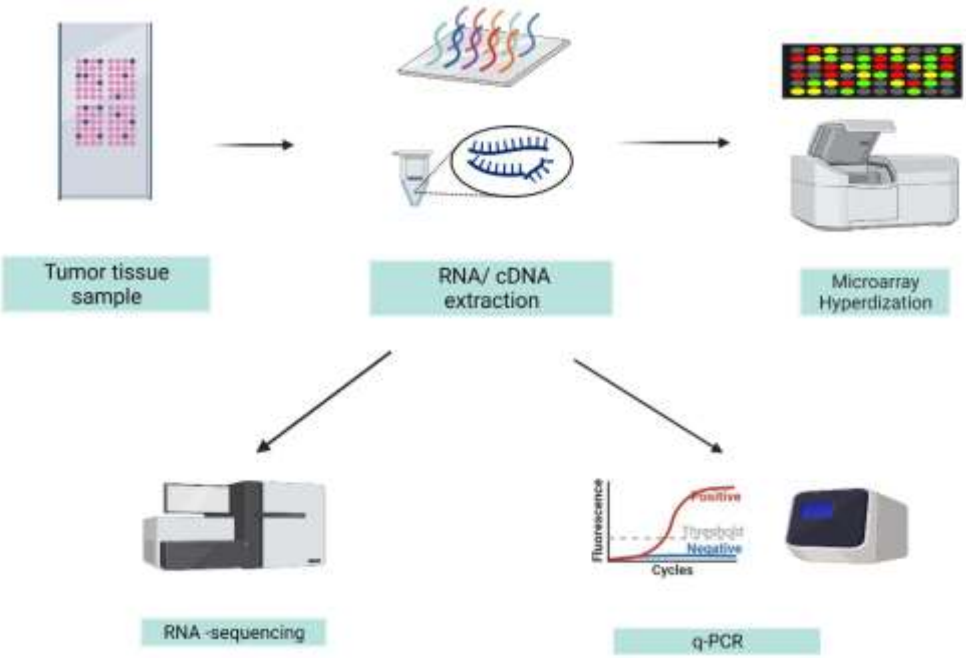


Figure 1.1: Overview of the steps for gene expression analysis

q- PCR, microarray and RNA-sequencing are the common techniques used for gene expression analysis

Source: created with BioRender.com

Subsequently, high-throughput methods such as DNA microarray technology have been used to analyse gene expression patterns on a large scale (Bernard & Wittwer, 2002). In this technique, a collection of DNA fragments (probes) are fixed on a solid surface (glass slide) in an ordered manner. Then each probe is specifically hybridized with targeted genes derived from a biological sample. The signals produced from this interaction are quantified and normalized to be used as indications for gene expression. The technique is beneficial in research settings, providing a way to discover novel cancer molecular subclasses (class discovery), compare different classes parallel to each other (class comparison), and identify relationships and predictions of therapeutic responses (class prediction).

However, there are some challenges in using microarray technology, of which the foremost is the need for more laboratory standardization. Each laboratory has its handling and analysis procedures, which results in bias that limits the use of the data in consecutive retrospective studies. Different microarray platforms can be used for gene expression analysis, employing other protocols for preparing, synthesizing, and annotating probes and their hybridization. Consequently, it is difficult to compare data generated from different platforms, and it is difficult to merge data sets. Researchers have tried to overcome the issues associated with platform variations by performing careful data pre-processing and normalization (Tinker et al., 2006). In this study, we extend this by utilizing a parallel artificial Neural Network based data mining technique to enrich for biomarkers addressing a given question and then identify concordances between the enriched biomarker lists. Another common high- throughput technique that is also used for transcriptomic analysis of biological data is RNA sequencing. It involves extraction, fragmentation and conversation of total RNA from cells or tissue into complimentary DNA (cDNA). The cDNA is then sequenced using high-throughput technologies, such as Illumina systems. Which are then aligned to a reference genome to quantify gene expression levels. RNA sequencing provides more comprehensive and accurate transcriptomic data compared to microarrays hence it is gaining more popularity. There is also differences in data analysis pipelines between the two technologies. RNA-sequencing data analysis includes alignment, quantification, differential expression analysis, and the detection of alternative splicing and novel transcripts, whereas microarray data analysis typically entails normalization and statistical testing (Wang et al., 2009).

1.3. Machine learning approaches for data mining

1.3.1. General overview

The Application of molecular biology techniques generates massive data with little biological interpretation. A common challenge for researchers is translating results generated by large molecular data matrices into a better understanding of biological processes relating to phenotype. There is considerable demand for approaches to analysing such data more efficiently and effectively. Statistical and machine learning approaches are the leading computer science techniques germane to this field (Grześ & Kretowski, 2007). Machine learning methods can be used for feature selection and classification from large matrices, rendering them suitable for comprehensive analysis of gene expression data. These approaches can be generally classified as supervised (when the study is based on existing biological knowledge about the gene) or non-supervised (if no predefined knowledge is used and the analysis is completely based on the data pattern).

Moreover, approaches for modelling gene expression data are continuously evolving. Each technique has unique strengths and weaknesses, which can serve the best in a specific research situation. It is up to researchers to select the analytical tools most commensurate with their particular needs based on the experimental design. Some of these tools are suitable for the comparative detection of gene expression patterns across multiple assays to discern better their biological functions and regulation, which leads to a better understanding of the disease (Narrandes & Xu, 2018). Others are helpful in investigating the entire components of biological systems to identify driver genes with high functional impact. In some research settings, these approaches could be used together to enhance the results and provide more meaningful insights from the data (Quackenbush, 2001). Moreover, machine learning approaches have advantages over other statistical methods due to their ability to handle high dimensionality and non-linearity associated with complex data. Machine learning approaches also offer adaptability by providing the option of parameter adjustment, which enhances classification performance. These advantages make them more appropriate for the analysis of gene expression data (Wuest et al., 2016).

Figure 1.2 presents a general schematic overview of the use of machine learning for the analysis of gene expression data.

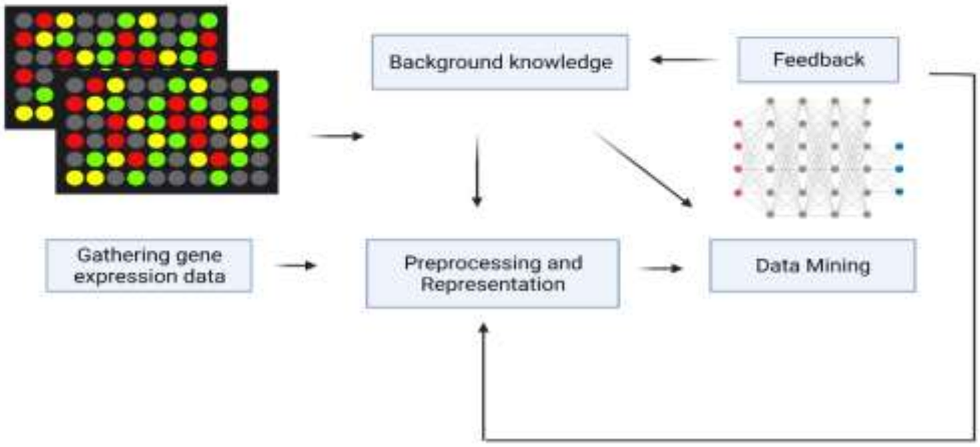

Background knowledge
Feedback
Gathering gene expression data
Preprocessing and Representation
Data Mining

Figure 1.2: Overview of machine learning environment

Source: adapted from Grzes and Kretowski (2007), modified using BioRender.com

1.3.2. Clustering approaches

The clustering of microarray data in biomedical research was pioneered by Michael Eisen et al. (1998). Clustering techniques can group genes based on their similarities in expression space, which can be visualized in a graph. They are a form of unsupervised machine learning. A cluster is represented by internal coherence and external isolation. The similarity between sets is defined using a distance measure, such as Euclidean, Cosine, Jaccard, or Edit distance. Clustering can be applied for multiple purposes, including identifying new disease subtypes and investigating mechanisms of gene regulation. By forming clusters, researchers can discover patterns in the data; however, these methods have limitations. They are very subjective as there are many algorithms for analysis and many ways to define the similarity that leads to different outcomes.

In addition, clustering is not suitable for prediction studies. In cluster analysis, distance measures are used to distinguish between classes that cannot reflect the influence of the relevant genes. Furthermore, clustering is unsuitable for comparison studies as it cannot provide valid statistical quantification of gene expression. Researchers using average fold change cannot determine the exact variability of gene expression across samples (Simon et al., 2003).The following subsections present some of the commonly used clustering approaches. This class also includes the neural network approaches which will be described in details in Chapter 2.

1.3.2.1. Hierarchical clustering

Hierarchical clustering approaches are a group of techniques widely used for microarray data analysis. The idea is to assign genes into clusters based on their expression. At each step, the two closest sets are identified and joined to produce a final tree called a dendogram, which is then used to define a meaningful biological pattern. There are two ways of constructing dendogram: bottom-up, and top-down.

Bottom-up dendogram construction (i.e., agglomerative clustering) assigns each gene to an individual cluster. The genes are agglomerated to produce small clusters, and the process is reiterated until one final cluster that includes all composite genes is produced. Figure 1.3 represents a final dendogram arising from hierarchical clustering. The distance between clusters is calculated based on pairwise dissimilarities using one of four methods: (1) singlelinkage, which measures the minimum distance dissimilarities between clusters; (2) complete linkage, which measures the maximum distance; (3) average linkage, which uses the average of all distances of the points between two clusters; and centroid linkage, which is based on measuring the distance between cluster centroids (Chipman & Tibshirani, 2006).

Top-down (divisive) clustering is less commonly used. It starts with one cluster and continues splicing into subgroups by identifying the greatest dissimilarities between the clusters (Alon et al., 1999). Hierarchy clustering approaches had complexity and timing issues, especially when the number of hierarchies increases. Because of the intricacy and timing concerns with hierarchy clustering methods, especially as the number of hierarchies grows, it takes more effort from the user to make accurate predictions (Rezende et al., 2022).

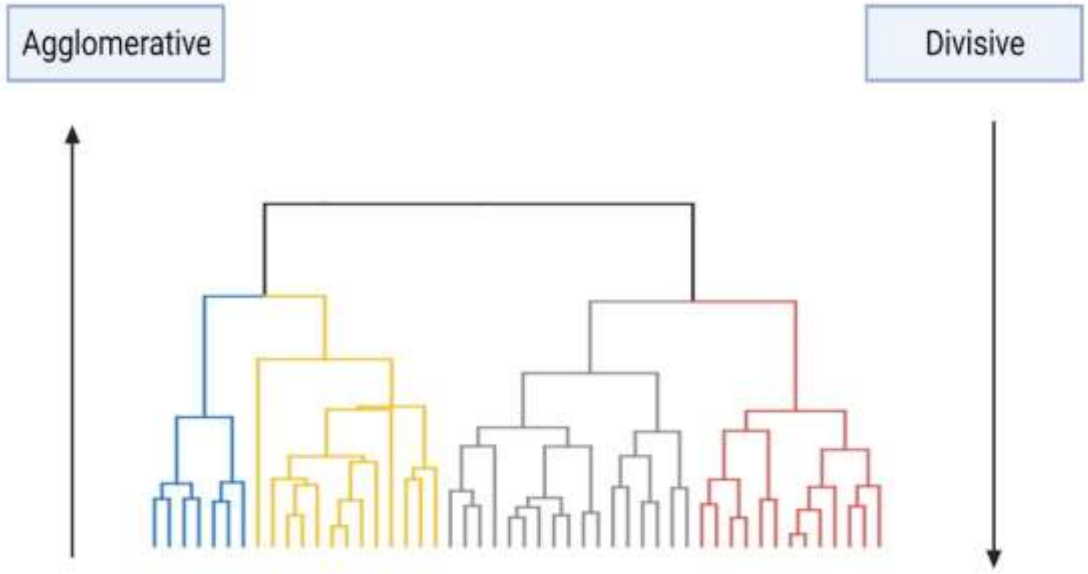


Figure 1.3: Dendogram presentation of hierarchical clustering

Source: adapted from Pirim et al. (2012)

1.3.2.2. K-means clustering

K-means clustering can be used as an alternative to hierarchical techniques in cases where the number of clusters is known *a priori*. This method's principle is assigning genes randomly into a predefined number of clusters (K). After that, a calculation of the average expression profile (centroid) is done for each group. Then genes are regrouped from one cluster to another based on their proximity to the available centroid. Calculating centroids and regrouping of the genes is performed iteratively until optimization or convergence is reached, which is a state of no further improvement of cluster composition (Madan Babu et al., 2004). Total squared Euclidean distance is used for the calculation of the centroid and the distance between genes in the same cluster. The method can help in building new classifications based on previous knowledge, such as the classification of patients with similar disease phenotypes and different clinical morphology based on the expression profiles. However, the method requires predetermination of the cluster numbers and can generate different results due to the initial random assignment of the genes (Xu & Wunsch, 2010). The requirement of pre-setting the value of K before running the algorithm represents a weakness of the K means clustering.

Although there are methods for setting it automatically, the majority of these are based on multiple random centroids initializations (Botía et al., 2017).

1.3.2.3. Self-organizing map (SOM)

Also known as Kohonen's self-organizing map, a group of nodes is created in this method, to which genes are assigned by proximity. The nodes are presented in a two-dimensional geometric space. Initially, a random gene is selected, and the nearest node (called a reference vector) is moved toward that gene; the other nodes are also adjusted based on how close they are to the selected gene. The process is repeated until no further adjustment in the positions of the nodes is possible. Then a final map of clusters represented by nodes and genes around them is produced. SOM has some advantages over K-means, including that it is flexible and more reliable. The number of final clusters is not necessarily equal to the starting one, as some of the nodes are without genes assigned to them and may end up being removed from the final map. However, SOM has some drawbacks. For example, you need to specify the number of clusters, which sometimes could be difficult, and it also requires similarity in behaviour between the nearby points to initiate the clusters (Madan Babu et al., 2004). Figure 1.4 displays the principles behind K-means clustering and SOM methods.

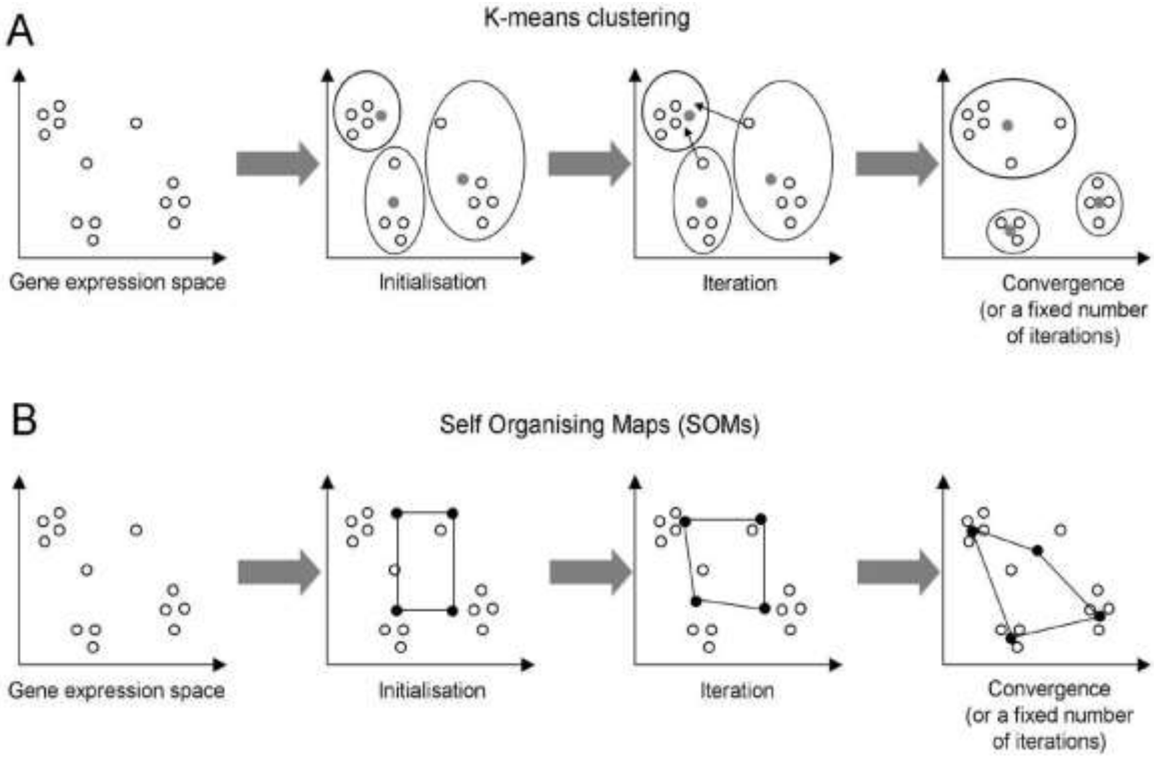


Figure 1.4: : K-means clustering and self-organizing maps (SOM)

Source: adapted from Babu (2004)

1.3.3. Principal component analysis (PCA)

PCA is a mathematical method used for data visualization and dimensionality reduction. It is also a form of unsupervised machine learning technique. Data variability is presented as an average set that summarizes the features of the data, and linear combinations of the original data are performed to produce a new set of variables (principal components) that can describe the data variability (Quackenbush, 2001). The most significant degree of data variability that can provide better separation of the data is named PCA1; the second presentation of data variability is named PCA2, and so on, until the maximum number of components is reached. The result is visualized in two- or three-dimensional plots that present the first few principal components. PCA can be used to explore the relationships between variables and to study the underlying processes in the data (Todorov et al., 2018). It is often used prior clustering techniques to determine the number of clusters. However, to enhance the quality of clustering, preliminary information about the data is still needed to choose the correct number of components (Yeung & Ruzzo, 2001).

1.3.4. Classification approaches

They were also known as supervised machine learning techniques. These methods use training data to recognize and characterize complex gene expression patterns. In these methods, models are first trained to distinguish the expression features of each class in the data and then assign each gene to its class, which can then be used for the classification of new genes that were previously unclassified. There are different clinical applications for classification approaches, including disease staging and stratification of patients to identify potential therapeutic responders. Classification techniques can also help in exploring new genes that are related to a known biological system, aiding in understanding mechanisms that lead to disease initiation and progression (Quackenbush, 2001). There are generally three broad types in this category; Support vector machines, tree-based approaches, which are described in the following subsections, and neural networks, which will be explored in chapter 2, since they are the primary analytical methods in this thesis.

1.3.4.1. Support vector machine (SVM)

SVM is a supervised method used for the classification of the data by a maximal distance hyperplane, which defines members from non-members of certain classes. It is a popular data mining tool with a good performance on data with multiple attributes, even if fewer samples are available for training (Bhaskar et al., 2014). In SVM, data is presented in a higher dimensional space (feature space), in which the distance between classes is measured using Kernel mathematical function. In some cases, misclassification could occur in SVM due to data noise; SVM addresses this issue using a soft margin that allows training errors (Ringnér et al., 2002). The method has been widely used for the analysis of gene expression data since an SVM could use previous biological knowledge from the training data to define the character of a given functional class and use this information to predict whether any additional genes could also belong to the class (Brown et al., 2000).

1.3.4.2. Tree based approaches

Tree-based tools are among the most popular machine learning tools applied in biomedical research because they are simple and easy to use, having good prediction performance with high-dimensional data. Decision tree models use tree-structured classifiers with the decision and leaf nodes. Although decision trees are mainly used for the prediction of outcomes based on specific data categories (i.e., classification trees), they can also be used where the data has continuous values (regression trees).

In decision nodes, there are two main branches that represent the outcomes based on known categories, while leaf nodes show the final classification or the value of the examples. The idea behind these methods is an iterative partitioning of the data based on the value of a selected example. They usually start by defining a root node with a known value and use it for continuous defining of a corresponding branch until reaching a leaf node that has the predicted value of the example, which can be a classification or regression. Decision trees are helpful in data analysis. However, they are non-robust and have a low accuracy compared to other supervised machine learning methods. Also, the topology of the trees is unstable, as a minimum change in the attributes can lead to a totally different result (Chen et al., 2011).

1.4. Systems biology and pathway analysis

1.4.1. General overview and operational definitions

Systems biology can be defined as “the study of complex interactions in biological systems and the emergent properties that arise from such interactions” (Du & Elemento, 2015). It encompasses multiple approaches aimed at exploring how biological entities interact and function within a defined system. Combining a detailed understanding of system components with a comprehensive analysis of the system in its greater context allows for the analysis and prediction of biological function (Kohl et al., 2000). Moreover, system biology analysis provides a useful way to address genome complexity in cancer. By modelling and integrating genomic data to view the full picture of how genes and pathways interact in cancer, system biology has a significant impact on the identification of novel properties (Werner et al., 2014).

Several cancer research studies implemented systems biology-based analysis to facilitate cancer treatment, including biomarkers identification for the detection of therapeutic response or optimization of treatment dose (Du & Elemento, 2015). Other studies used a pathway model for the representation of the biological processes in which genes and their products interact in an ordered manner to achieve certain biological functions (Mitrea et al., 2013). Biologists use the term "pathway" to provide a description of specific biological processes. Demir et al. (2010) defined a pathway as "a set of interactions between physical or genetic cell components, often describing a cause-and-effect or time-dependent process, which explains some observable biological function." The term "network" is also used in a technical sense to refer to integrative analyses of multiple datasets to gain insights into biological systems (Creixell et al., 2015).

1.4.2. Computational approaches for pathway analysis

A significant interest in the pathway and computational network analysis has emerged in cancer research. Pathway modelling is important because it reflects the biological relevance of genes under investigation and helps make predictions about cellular processes in health and disease. The power of pathway modelling relies on its ability to extract meaningful biological knowledge associated with a particular phenotype from a list of differentially expressed genes. Since genes that are differentially expressed often participate in common pathways, and the alterations observed either enhance or suppress the pathway activity, it is, therefore, crucial to identify pathways involved in cancer and detect their mechanism of alteration. This can give an indication about certain phenotypes, which could help in disease diagnosis and personalized treatment (Vaske et al., 2010).

Hence, several computational approaches have been innovated with the potential to improve the investigation and representation of the entire components of pathways and to identify driver genes with high functional impacts. Moreover, computational models provide essential insights into pathways that drive disease progression, and they can be used to test hypotheses and make predictions. Khatri and Drăghici (2012) provided a general overview of the existing pathway analysis methods and generally grouped the methods by the type of analysis into three major classes: over-representation analysis, functional class scoring, and pathway topology-based methods. The following subsections explain these classes, as depicted in Figure 1.5.

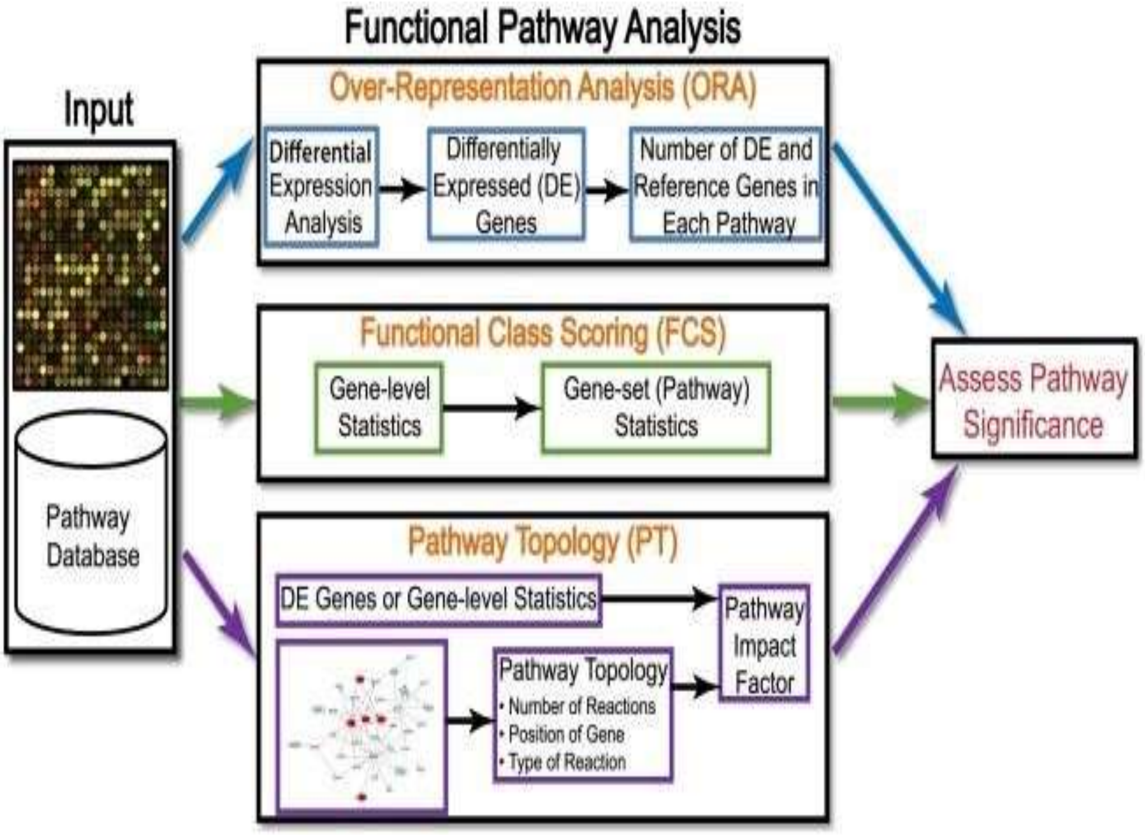

Functional Pathway Analysis
Input
Pathway Database
Over-Representation Analysis (ORA)
Differential Expression Analysis
Differentially Expressed (DE) Genes
Number of DE and Reference Genes in Each Pathway
Functional Class Scoring (FCS)
Gene-level Statistics
Gene-set (Pathway) Statistics
Pathway Topology (PT)
DE Genes or Gene-level Statistics
Pathway Topology
• Number of Reactions
• Position of Gene
• Type of Reaction
Pathway Impact Factor
Assess Pathway Significance

Figure 1.5: Overview of common pathway analysis approaches

Expression data used as input for all pathway analysis methods. ORA methods use differential gene expression analysis, while FCS use a whole data matrix, and PT-based methods consider the type and the number of genetic interactions

Source: adapted from Khatri and Drăghici (2012)

1.4.2.1. Over-representation analysis (ORA)

This class aims to evaluate a set of genes within a particular pathway found among a set of differentially expressed genes. They were created primarily for the computational detection of somatic mutations, facilitating the comparison of a set of mutated genes to a known pathway from databases to identify overlap using statistical testing. If the overlap is statistically significant, the list can be considered enriched in relation to the prospective pathway. These methods also use statistical measurements to assess random errors, including Fischer's exact and hypergeometric tests (Dimitrakopoulos & Beerenwinkel, 2017).

Gene ontology (GO) uses an over-representation statistical approach (ORA) to categorize differentially expressed genes to certain functional classes (GO categories) by comparing the number of genes found in each category of interest with the number that may occur by chance; the class is considered to be significant if the number reported is substantially different from the one that may be assumed to occur randomly. This approach can provide general biological themes in the selected genes, but detailed analysis is not offered by this method, as this would result in a vast number of gene categories, which would be untenable to collate and analyse. Furthermore, putting genes into selected classes could cause important gene patterns to be lost. Moreover, the decision to put genes in a certain class is based on a certain threshold, which means that the class could change if the threshold itself changes. There are many tools that use GO methods for pathway analysis, including DAVID, GoMiner, and GOToolBox.

A broad review of ORA methods reveals shared limitations, including the use of a threshold method to define the most significant genes. With this method, some information about the marginal and less significant genes could be missed, and ORA methods need to consider overlapping pathways or enable the evaluation of each gene individually. (e.g., to consider the interactions between different candidates within a pathway).

1.4.2.2. Functional class scoring (FCS)

This class analyses a whole data matrix as an input and assumes that coordinated weaker genes may also have significant impacts on pathways. FCS performs computational analysis at the gene and pathway levels, using different statistical methods, including analysis of variance (ANOVA), T-test, and Z-score. The most commonly used approach for the analysis of gene expression data is gene set enrichment analysis (GSEA), which is based on a functional class scoring (FCS) statistical approach. FCS considers the functional relation between the genes by including all gene expression values. Subramanian et al. (2005) used GSEA to identify over and under-expressed genes in a particular dataset in comparison to a predefined gene list. The list was already linked to a certain biologic pathway based on previous knowledge.

The method was further refined to create a software package that can be run as a desktop application, and it has subsequently been widely used as a common tool for pathway enrichment analysis. It has been valuable for the interpretation of large volumes of biological data, but GSEA uses FCS analysis such that it treats genes that have the same rank equally, even if there is a considerable variation in their expression. GSEA also has the limitation of ORA methods in being based on curated pathways identified from previous literature, thus making it impossible to predict novel pathways. This is because it is tied to previous knowledge of cancer pathways, disregarding the crosstalk between different pathways by considering them as specific groups (Subramanian et al., 2005).

1.4.2.3. Pathway topology (PT) methods

PT methods provide details about the interaction between gene products by answering *how* and *where* questions to provide information about the nature and the position of the genetic interactions. PT-based methods use the same framework as FCS, but they utilize pathway topology to perform the statistical calculation at the gene level. Examples of these methods are Reactome, Panther, and Kyoto Encyclopedia of Genes and Genomes (KEGG). These were developed to overcome the primary challenge of pathway analysis by providing repositories to collect and present the complex mechanisms of the pathways and to facilitate the analysis and modelling of large biological systems. In these approaches, genes are represented as nodes, and the interactions between them are represented as edges. These approaches are based on known knowledge from literature, and they use available databases to identify significant pathways of a given gene expression data. Moreover, an impact analysis model proposed by Draghici et al. (2007) based on the calculation of an impact factor can identify pathways that are significantly changed in a certain condition. It has the advantage of including some biologically meaningful changes on a given pathway, such as the magnitude and position of differentially expressed genes within a pathway. The method has been developed and published as a web-based tool (PathwayExpress).

1.4.3. Graphical methods

Several algorithms have been implemented for the analysis of cellular networks using graph methods for data integration and network modelling, whereby cellular components are presented as nodes, and the interactions between them are presented by edges. For instance, gene regulation networks can be graphically modelled with transcriptional components being presented as source nodes and regulated factors as sink nodes. This type of graphical representation helps in understanding the topology and function of cellular networks and allows for the prediction of new biological hypotheses, such as exploring new interactions that can be tested using laboratory experiments (Aittokallio & Schwikowski, 2006). Moreover, mathematical models are used for iterative reconstruction of the network, which provides a more detailed and accurate prediction of the network properties (Papin et al., 2005). Databases such as STRING (Franceschini et al., 2012) and GeneMANIA (Warde-Farley et al., 20110) are commonly involved with this type of analysis.

Moreover, Leiserson et al. (2015) developed HotNet2, a network-based algorithm used to analyze data from The Cancer Genome Atlas (TCGA) of 12 cancer types. In this method, a directed network heat diffusion model is used, in which genes are presented by nodes and genetic interactions by edges. A heat score is assigned to each gene according to the frequency of alteration, and heat diffuses to other nodes in the networks across the edges. Thus nodes that receive significant amounts of heat based on (statistical modelling) are reported. The method identifies genetic combinations and provides new insights into the interactions between genes in well-known cancer signalling pathways on a large scale.

However, the presence of highly mutated and highly connected genes in the network generates extremely hot nodes, affecting nearby nodes and leading to false positive results, limiting the accurate detection of rare mutations.

Several software tools have been implemented for graphical network visualization, including Cytoscape, MetaCore software from Thomson Reuters, and Ingenuity Pathway Analysis. These tools allow for construction and visualization of multiple pathways and are helpful in the interpretation of biologically significant results and in drawing conclusions. However, there are some challenges associated with this analysis. The method uses linear modelling, which cannot cope with the multi-directionality of real genetic interactions. Data from different cancer types are also merged, which limits the specificity of the results (Leiserson et al., 2015).

1.4.4. Biological knowledge-based approaches

PathOlogist approach uses structural pathway knowledge to quantify the nature of interactions in a pathway. The method estimates the probability and constancy of interaction in three steps: (1) estimating functional activity/inactive genes based on their mRNA expression values using clustering algorithms; (2) determining interaction activity; and (3) estimating the overall average of active interactions and using this as an indication for the activity and consistency of the whole pathway (Efroni et al., 2007). Moreover, Tarca et al. (2009) developed Signalling Pathway Impact Analysis (SPIA) to capture the impact of gene expression changes on a pathway, addressing issues in methods known for overrepresentation, such as GSEA, which identify the significance of differentially expressed genes in a pathway.

SPIA added a perturbation analysis, which considers the impact of differentially expressed genes in the pathway by considering their position and assuming that expression changes in a rooted gene (that influence several interactions) within a pathway could be highly significant than changes in a leaf gene (which has no influence on other interactions). Vaske et al. (2010) proposed a method called Pathway Recognition Algorithm using Data Integration on Genomic Models (PARADIGM) to infer patient-specific pathway activities by integrating cancer genomewide data obtained through multiple multi-omics technologies into a pathway framework. The model utilizes curated interactions from pathway databases and converts them into a graphical factor model, integrating information that describes states of cell components (e.g., mRNA level) with known interactions.

1.5. The TP53 pathway

1.5.1. Structure and function

The human TP53 gene is located at the short arm of chromosome 17(17p13) and spans about 20KB of DNA. The TP53 protein comprises 393 amino acids and four functioning domains (May & May, 1999). The TP53 protein was originally identified by Lane and Crawford (1979), while conducting research on the effect of a Simian virus 40 (SV40) antigen on tumour bearing host cells. They observed a cellular protein with an apparent molecular mass of 53KDa, named TP53, and they concluded that TP53 interacted with SV40 antigen (Lane & Crawford, 1979). Subsequent studies indicated that the TP53 protein is a growth regulatory molecule and a cell cycle dependent protein (Louis et al., 1988; Milner, 1984; Reich et al., 1983).

TP53 acts as a transcription factor to regulate multiple downstream genes. The protein is highly expressed in response to several types of stress signals, such as oncogene activation, DNA damage, and hypoxia. Harris and Levine (2005) defined the active TP53 gene by its ability to bind specifically to the promoter region of the targeted genes, which results in increasing the half-life of the protein from around 30 to 150 min on average. Activation leads to the expression of mediators and core regulatory genes to produce a variety of cellular responses, including cell-cycle arrest, DNA repair, and apoptosis. Each of these responses is generated through a specific signalling pathway. Figure 1.6 shows these functional pathways and illustrates the interpretation of the signals through different downstream regulatory genes.

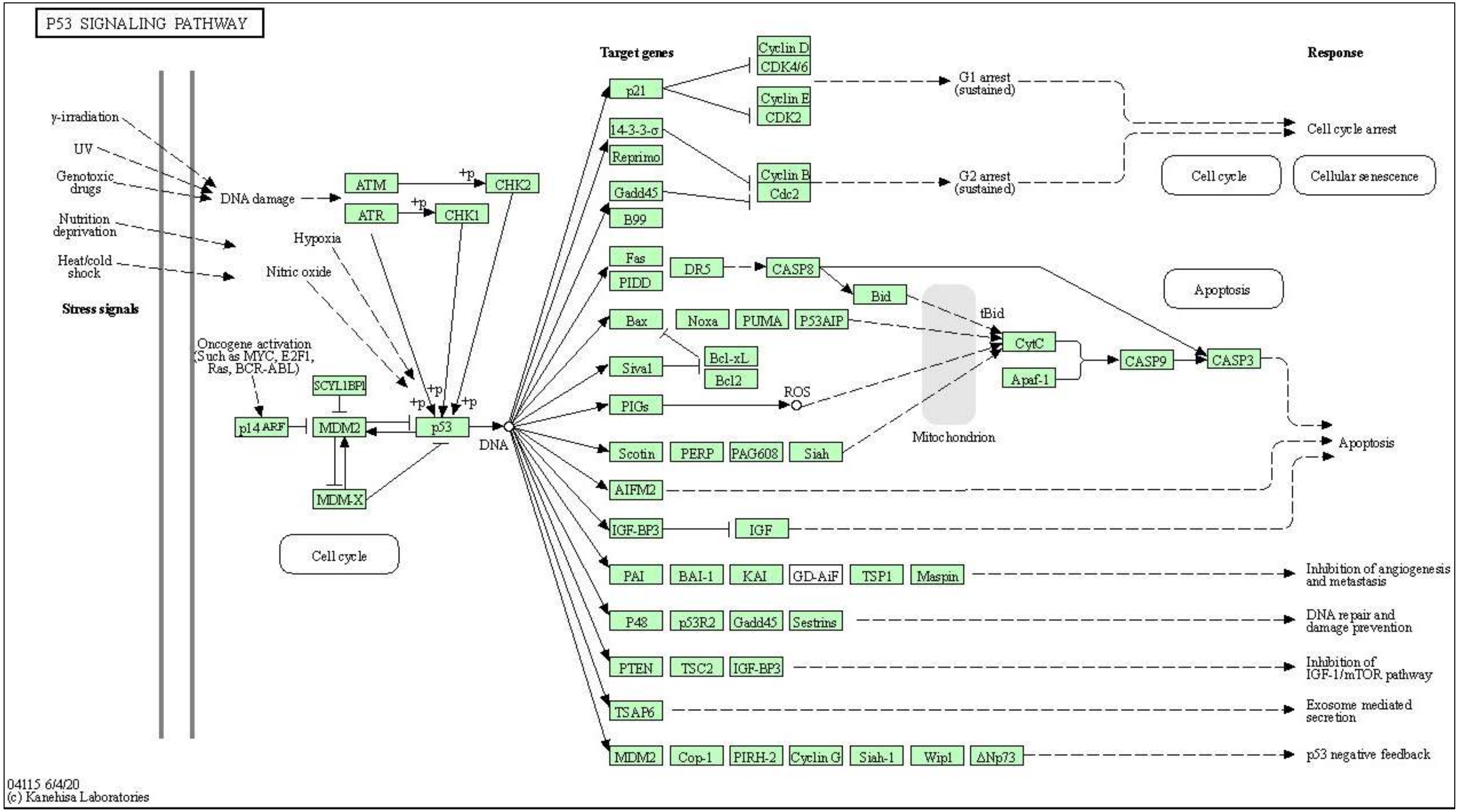


Figure 1.6: Schematic representation of TP53 pathway

Multiple stress signals activated downstream mediators and core regulatory genes to produce various cellular responses. Source: adapted from KEGG database.

There are factors which determine which of these responses will be chosen, including cell type, the nature of the stress signal, and the extracellular environment proteins. For example, excessive cell division triggers activation of the *ATM* and the*TP53* genes to initiate cellular senescence response. Inactivation of the *TP53* gene reverses this mechanism, and allows more cell division, which result in shortening of the chromosomal telomers and a massive cell death (Vaziri, 1997). The best characterized functions of the *TP53* gene are cell cycle arrest and apoptosis, as discussed in the following sections.

1.5.2. TP53-mediated cell cycle arrest.

Kuerbitz et al. (1992) reported the Wild-type *TP53* gene (*WTTP53*) as an important determinant of cell-cycle arrest following exposure to ionizing radiation (IR). The study indicated the effect of the *TP53* gene alteration on human cell cycle progression. Although the study was primarily designed to study the *WTTP53* gene, the authors decided to turn the focus of the study to demonstrate the inhibitory effect of the Mutant *TP53* gene on the cell cycle in different cell lines. Since the original experiment suffered from growth and selectivity disadvantages occurring from the endogenous *TP53* gene during the transfection of the *WTTP53* into cell lines, Kastan et al. (1992) proposed a G1-cell cycle arrest pathway upon exposure to IR. The pathway involves the activation of ataxia-telangiectasia (*AT*) genes, the product of which leads to an increased *TP53* level. The study also identified *GADD45* as a downstream contributor to the pathway and highlighted the importance of *TP53* for the activation of *GADD45*. However, the mechanism by which *GADD45* induces cell-cycle arrest was not indicated.

El-Deiry et al. (1993) introduced *P21/WAF1* as an important downstream target of the *TP53* pathway and a potential mediator of the *TP53*-mediated growth inhibitory response. Although they did not clearly identify the role of *WAF1* and the mechanism that led to growth inhibition, they used data from a contemporaneous study by Harper et al. (1993) to suggest that the *WAF1* protein was coupled with the product of another gene, named *CIP1*, and this interaction blocked cell cycle progression by inhibiting cyclin-dependent kinase activity.

Hermeking et al. (1997) demonstrated induction of the *14-3-3 σ* gene as a result of the *TP53* activation following the treatment of colorectal cancer cell lines by DNA-damaging agents. The study showed that the *14-3-3 σ* mediated a coordinated cell-cycle arrest by blocking the transition of the cell from G2 to the M phase. The mechanism involves binding of the *14-3-3* protein to the *CDC25C*, and inactivation of an important cell cycle gene, cyclin-dependent kinase (*CDC2*), which is required for the entry of the cell into mitosis. Figure 1.7 illustrates the model of TP53-mediated cell-cycle arrest.

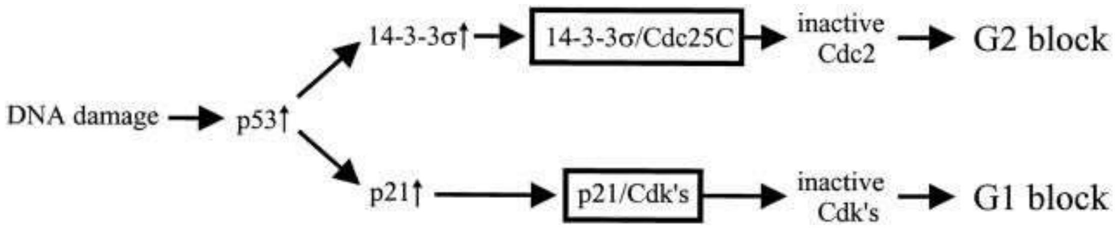


Figure 1.7: Model of TP53-mediated cell cycle arrest

Growth inhibitory signal induced by the TP53 in response to DNA damage leads to activation of downstream genes that cause inactivation of cyclins and block the cell cycle

Source: adapted from Hermeking et al. (1997)

1.5.3. The TP53-mediated apoptotic response.

TP53 promotes apoptosis in response to different stimuli, with mechanisms involving the activation of different targets that eventually lead to programmed cell death. There are two main apoptotic pathways in mammalian cells. The first is the *BCL2*-regulated pathway, also called the intrinsic pathway. In this pathway, cell death is initiated by the activation of proapoptotic members (*NOXA, PUMA, PIM*), which exerts an inhibitory effect on *BCL2*-proteins (*BCL-XL, BCL2-MCL-1*). This leads to the activation of *BAX* and *BAK* and the generation of apoptotic signals, with the consequential release of mitochondrial cytochrome c, which binds to the apoptotic protease activation factor *(APAF1).* Many *APAF1* molecules are aggregated together to form an apoptosome, which recruits and activates the caspase-9 enzyme. This enzyme is responsible for the cleavage and activation of a series of other caspase enzymes, which collectively induce cell death. The second pathway is the death receptor pathway, also called the extrinsic pathway, which involves activation and cleavage of caspase-8 by *FADD* adaptor protein, leading to a consecutive activation of caspase-3 and -7, which eventually causes apoptosis (Aubrey et al., 2017).

1.5.4. TP53 positive and negative feedback loops.

There are gene products within the *TP53* network responsible for autoregulation of *TP53* activity and pathway communication with other signalling pathways. These proteins can turn the *TP53* protein on or off. Most of them function in a series of feedback loops involving the *MDM2* protein (which effectively turns the *TP53* protein on or off). Among these, three increase *TP53* activity *(p14/19ARF, PTEN-AKT, and Rb)*, and seven inhibit it *(MDM2, TP73, Cop1, Pirh2, Wip1, Cycling, and Siah1).* Another function of these proteins is to connect the *TP53* pathway to the other pathways and regulate the signals for cellular growth. For instance, Wip1 protein connects the *TP53* to the Ras/Raf/Mek/Erk pathway through a negative feedback loop involving P38 MAP kinase (Takekawa et al., 2000); and Siah1 connects the TP53 to the Wntbeta-catenin-APC pathway (Harris & Levine, 2005).

MDM2 proteins play a central role in the regulation of the *TP53* activity. They are frequently detected in human tumours harbouring Wild-type (but not Mutant) TP53 (Oliner et al., 1992). The relationship between *MDM2* and *TP53* is bidirectional: *TP53* acts as a transcriptional activator for the *MDM2* gene, while *MDM2* serves as a negative regulator of *TP53*. The formation of a *TP53-MDM2* complex leads to ubiquitination and degradation of *TP53*. These bidirectional relationships are important for maintaining the balance of the two proteins and limiting the duration of the *TP53* activity upon stimulation. Moreover, there are mediators which enhance the degradation of the *TP53* by *MDM2*. Among them is *Wip1* (Wild-type TP53 induced phosphatase 1). This gene function as stabilizer for *MDM2* and enhancer for the *TP53/MDM2* ubiquitination (Lu et al., 2008).

Furthermore, *MDM2* promotes cellular growth through a mechanism involving phosphatidylinositol 3-Kinase (P13-kinase) pathway. The signal from this pathway enhances the movement of *MDM2* from the cytoplasm to the nucleus, where it binds and inhibits the function of the *TP53.* This explains how mitogens could mediate cellular growth by modulating the *TP53* function and highlights the possibility of regulating the *TP53* gene by targeting components of PI3K/AKt pathway (Mayo & Donner, 2001).

Upon oncogenic stimulation, *MDM2-TP53* complex is negatively regulated by gene named alternative reading frame *(ARF)* gene. The inhibition of *MDM2* by *ARF* leads to the induction of the *TP53*, which exerts a major function as an inhibitor of abnormal growth. The *ARF* stimulatory effect on the *TP53* needs further investigation (Shi & Gu, 2012). Another inhibitor of *MDM2* activity is *14-3-3 sigma*, which exerts its effect by blocking *MDM2-TP53* ubiquitination and promoting the stabilization of the *TP53* (Yang et al., 2003).

1.6. TP53 pathway in cancer

1.6.1. Role of TP53 in tumour suppression

The TP53 protein has a crucial role in tumour suppression. Its ability to mediate apoptosis plays a significant role in tumour clearance. It acts as a sensor for a wide variety of oncogenic stress signals, to inhibit tumour development and to limit the propagation of cells under stress

(Vousden & Prives, 2009). For low-level stress, *TP53* engages DNA repair and a temporary program of cell-cycle arrest to allow cells to pause and repair the damage; conversely, in response to more potent stimulus, *TP53* induces cellular senescence. In other situations where the stress signals are severe or sustained, *TP53* responds by activating components of the death pathways, including *BAX, NOXA, FAS, and PUMA,* which eventually lead to irreversible apoptosis or senescence.

Another form of the tumour-suppressive activity of *TP53* occurs through the inhibition of glycolysis and induction of oxidative phosphorylation in response to various metabolic stresses, including hypoxia and nutrient depletion. *TP53* also limits cancer formation through autophagy or “self-eating”, which leads to cell death by activation of target genes, such as *SESN1/2* and *DRAM. TP53* has an antioxidant function that protects cells from the damaging level of oxygen species (Vousden & Ryan, 2009). Moreover, *TP53* plays a role in the inhibition of tumour angiogenesis through upregulation of angiogenesis inhibitors, downregulation of proangiogenic genes, and inhibition of the hypoxia-sensing system (Teodoro et al., 2007). TP53 can exert its effect upon tumour stromal tissue to inhibit tumour growth and metastasis by increasing the ability of stromal fibroblast to secret tumour inhibitory factors and suppressing the production of tumour-promoting agents (Bar et al., 2009).

Moreover, among TP53-family proteins, the *TP73* and the *TP63* genes are also involved in tumour suppression, and actively participate in the resulting cellular output. Through interactions with common and specific regulators, these family members function to govern apoptosis and cell cycle arrest during stress (Collavin et al., 2010). The *TP73* and the *TP63* proteins have remarkable structural and functional similarities, but each has some unique specializations. It has been proposed that they could be required for stable binding between the *TP53* and its targets, forming a large transcriptional complex holding all three proteins (Urist & Prives, 2002).

1.6.2. TP53 mutation in cancer

Inactivation and somatic mutation of the *TP53* is ordinary in human cancers. The gene is mutated in 50% of human tumours, making it the most frequently altered gene in human cancers (Kandoth et al., 2013). Missense mutations are the most common type of *TP53* mutation, occurring as point mutations in the central domain of the protein, and leading to amino acid substitutions and the formation of the Mutant *TP53* protein. The latter is more stable than the Wild-type TP53, and is often present at a high level in cancers (Vousden & Lu, 2002). Aberrant *TP53* loses its ability to suppress tumours and gain new functions that promote tumourigenesis, such as increasing cellular proliferation, evading apoptosis, and therapy resistance. Figure 1.8 demonstrates the oncogenic properties of the Mutant *TP53* and its underlying mechanism (Brosh & Rotter, 2009).

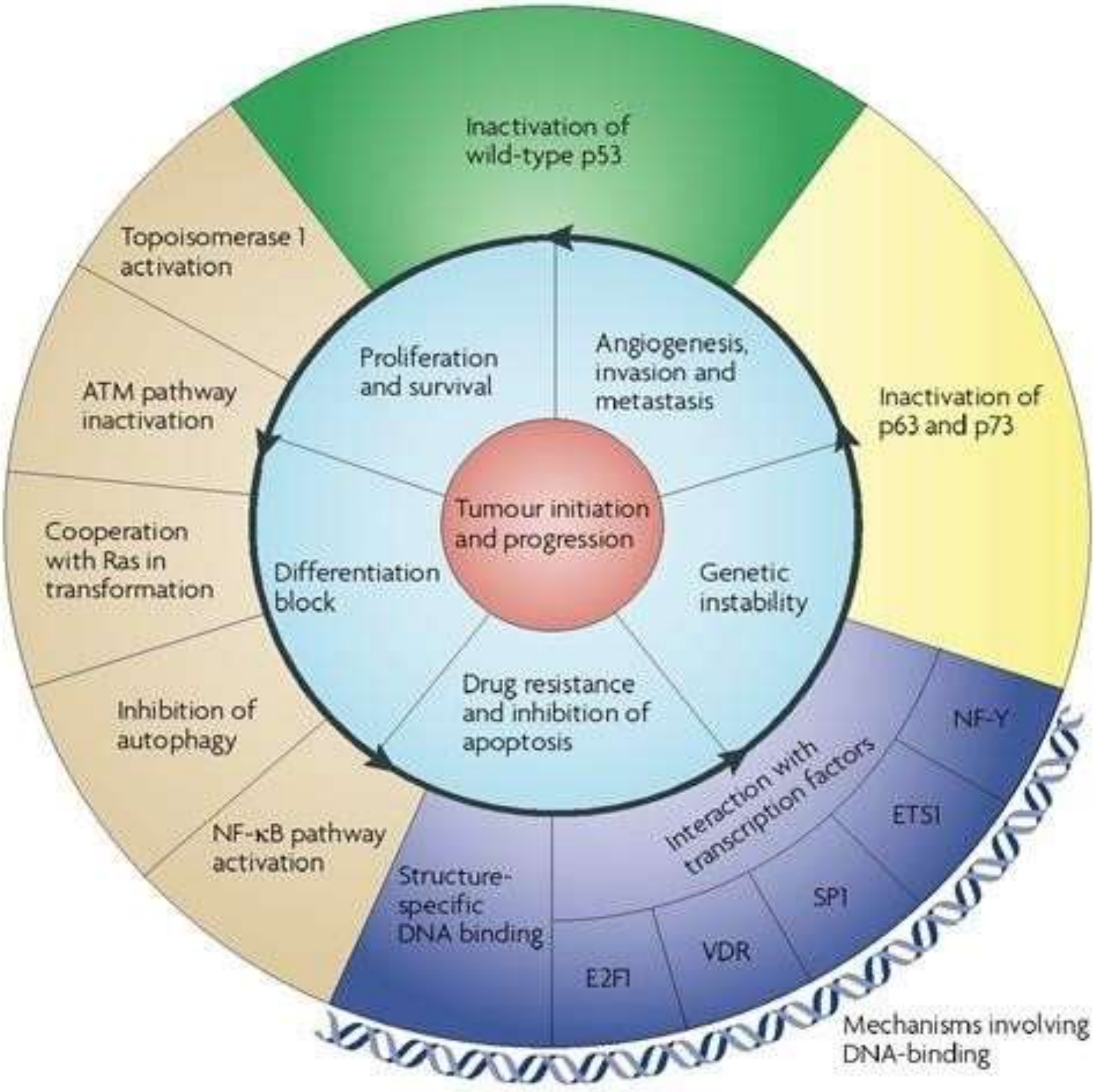

Inactivation of wild-type p53
Topoisomerase 1 activation
ATM pathway inactivation
Cooperation with Ras in transformation
Inhibition of autophagy
NF-κB pathway activation
Structure-specific DNA binding
Interaction with transcription factors
E2F1
VDR
SP1
ETS1
NF-Y
Inactivation of p63 and p73
Proliferation and survival
Angiogenesis, invasion and metastasis
Tumour initiation and progression
Differentiation block
Genetic instability
Drug resistance and inhibition of apoptosis
Mechanisms involving DNA-binding

Figure 1.8: Oncogenic properties of Mutant TP53

The oncogenic phenotypes of the Mutant TP53 is represented in the inner blue circle, while the outer circle indicates mechanistic properties for each phenotype listed in the inner circle

Source: adapted from Brosh and Rotter (2009)

Moreover, the mutations in the *TP53* have been linked to the gene expression patterns in human tumours, which can identify signatures that can be used as specific indicators for clinical outcomes in cancer. For instance, breast cancer harbouring *TP53* mutation had a specific gene expression pattern, which might be used as a survival marker for breast cancer patients. This association is strongly linked to certain classes, including the Basal-like molecular and ERBB2 amplification subgroup of breast cancer, wherein the *TP53* mutation occurs as an early event in tumorigenesis. Patients under these classes experience a shorter survival rate compared to other disease classes. The detection of the *TP53* mutation, together with specific gene expression patterns, may aid in distinguishing those patients at a higher risk of mortality (Langerød et al.,2007).

Abdelfatah et al. (2010) proposed the combined use of certain gene candidates from the *TP53* pathway *(MDM2/MDM4/BCL2 and P21)* as prognostic markers for breast cancer. The study used the protein expression of these candidates for the assessment of breast cancer patients. Two main classes were obtained: the high-risk group, which has a good prognosis and a favourable clinical outcome, and the low-risk group, with a poor clinical outcome and shorter survival timing. However, the association of the *TP53* mutational status with the clinical outcome in different types of tumours remains controversial. The majority (65-90%) of studies have linked *TP53* mutation to poor prognosis in colorectal, breast, bladder, and haematological malignancies; conversely, in lung, ovarian, and brain cancers the pattern is different, as half of the studies reported no association between the *TP53* mutational status and the clinical proprietaries. In addition, association with good prognosis is also noted in some cases (Brosh & Rotter, 2009).

Mutant *TP53* has become an attractive target in the cancer therapeutic era, and there is great interest in the selectivity between tumour and normal cells, as it increases the sensitivity of tumour cells toward therapy. Several therapeutic strategies have been developed to restore the function of Wild-type *TP53*. For example, the use of *MDM2* inhibitors can inhibit the degradation of *TP53* and enhance tumour regression by promoting cell death. However, as the *TP53* function can act as a guardian and survival enhancer for cancer cells, restoration of the Wild-type *TP53* is not always effective in cancer treatment. Indeed, in some cancer cases, this strategy can protect cancer from a certain type of cytotoxic drugs and is associated with poor response to treatment (Mandinova & Lee, 2011).

1.7. Computational modelling of the TP53 pathway

Several computational models have been innovated to study the TP53 network. These models refine knowledge and provide deep insight into the structure, mechanics, energy, and dynamics of both Wild- and Mutant-type TP53 in relation to other members of the TP53 pathway (Tan et al., 2019). Although it seems challenging to gain comprehensive insights into the TP53 network, various approaches have been developed and applied for network analysis in this context, which can generally be classified into interaction and mathematical models, as described in the following subsections.

1.7.1. Interaction models

Tuncbag et al. (2009) constructed a PPI network for hub proteins related to the TP53 pathway. The method used for this analysis is named Protein Interactions by Structural Matching system (PRISM algorithm). It was used to predict structural similarity and potential interactions that can occur within the pathway simultaneously. The authors presented the concept of integrating time into the interaction network and assumed that some genes with different binding sites could interact contemporaneously with the *TP53* gene while others that have similar binding sites could not. The method included structural information about the network and was useful in assessing part of the pathway functionality. However, PRISM algorithm uses data from the protein data bank, which means it does not consider the possibility of novel or indirect interactions. Csikász-Nagy et al. (2006) proposed a protein interaction network to model the activity of cyclin-dependant kinase in eukaryotic cells and the proteins that regulate them. The model presents a primary understanding of the cell cycle network across species.

Toettcher et al. (2009) combined a computational network with an experimental study to determine distinct mechanisms that mediate cell cycle arrest and re-entry in response to damage. The study found that specific mechanisms act to achieve arrest after damage and to stop improper cell cycle re-entry. However, the model was originally built to match data from yeast to mammals and was adapted to study cell cycle arrest in humans. Villaamil et al. (2011) identified networks related to the TP53 pathway in renal cell carcinoma using the STRING database and MeV bioinformatics tool. They modelled protein interaction and further validated the results with immunohistochemistry protein expression profile, and the results indicated two protein networks: one involved in angiogenesis pathway and the other indicates a negative association between *TP53* and glucose transporter type 4 (Glu4). However, the method was developed for a particular cancer type, which limits its generalization to other cancers. Although the results of this study were interesting, further validation using larger sample sizes and different sources could add value, as it only based on a sample of 80 patients from a single data source.

1.7.2. Mathematical models

Mathematical models provide logical representations of biological systems, allowing scientists to test hypotheses, make predictions, and gain new insights into biological processes

(Gatenby, 2012). Several mathematical models have been implemented to explore the *TP53* pathway, which provides a better understanding of the functional and dynamic properties of the *TP53* gene and its impacts. The first set of models was focused on the negative feedback loop between the *TP53* and the *MDM2* genes. Lev Bar-Or et al. (2000) developed a kinetic model based on the Ordinary Differential Equation model (ODE) to model the cellular concentration of the *TP53* and the *MDM2* genes. ODE has been used to study the rate of change of certain proteins within the *TP53* network with respect to time, and it can predict how the *TP53* dynamics influence the decision of cell survival and death. The partial differential equation has been used to understand special patterns of the TP53 gene (Kim et al., 2019). Consequently, other proteins, including *ATM/WIP1* and *P21* were also investigated using ODE models.

Sun et al. (2011) constructed a mathematical model to detect and characterize basal *TP53* pulses under stressed and unstressed conditions, thereby indicating the tolerant and sensitive nature of the *TP53* system. However, not all interactions were considered in this mathematical model, and it was integrated with experimental studies to control the dynamic behaviour of the *TP53* during stress conditions. Purvis et al. (2012) used computational model to show the possibility of controlling cellular fate by adding a timed drug that alters *TP53* pulses, leading to the expression of different sets of downstream elements. The study identified protein dynamics as an essential influencer of cellular fate. Cells with pulsing *TP53* dynamics recovered from DNA damage, while those with sustained *TP53* levels underwent senescence.

Purvis et al. (2012) also proposed a treatment strategy based on the induction of *MDM2* inhibition to alter the *TP53* dynamics from pulsed to a sustained level. Tian et al. (2017) used a dynamic network to unravel tumour-suppressive mechanisms in response to mitogenic and oncogenic signals, and the network described cell fate decisions with a focus on *ARF* as a major outcome of oncogenic signalling. Moreover, other mathematical models have been developed to investigate the *TP53* system in metabolism. A computational model was built to assess the cell fat decision by comparing signalling and regulatory network in autophagy and apoptosis, and the model predicted *TP53* as a regulator of cell fate transition from autophagy and apoptosis (Liu et al., 2017).

1.8. Challenges in pathway analysis

Draghici et al. (2019) undertook a comparative review of 13 widely used pathway analysis approaches. Although methods that consider the description of the pathways perform better than those based on a list of differentially expressed genes, the results indicate that no method has superior performance to others. In fact, most pathway databases and software analysis tools use curated pathways from the literature to turn gene expression lists into functional categories, which represents a major drawback. This is because the databases are incomplete, and most of them use manual curators to review existing knowledge from literature, which causes delays in the curation process.

Also, some of the approaches were built on linear statistical models, but in the field of cancer research, molecular biology data are usually non-linear and contain noise that needs careful consideration upon analysis. Moreover, existing approaches use electronic annotation without a system for error detection, which leads to the generation of inaccurate information (Khatri & Drăghici, 2005). It is, therefore, imperative to devise and use methods that can cope with these challenges. ANN tools can cope with noise and high dimensionality associated with molecular biology data and have a system for error detection and continuous adjustment of the results, thereby generating more accurate information.

1.9. Project aims

As described in section 1.6, the TP53 pathway has been known for its crucial role in cancer; it activates regulators that suppress the tumours via different cellular responses. Also, the TP53 mutations are present in about half of all tumours. In the remaining half, the entire pathway is disrupted due to dysregulation in other pathway components (Huang, 2021). The knowledge about these components and the interaction between them has been characterized and computerized in the database. However, this knowledge is based on the existing experimental findings in the literature, which means some information could be lacking. For this, we investigated the whole TP53 network to identify new drivers that could be added to the pathway. This aim will be achieved by exploring existing and new features of the TP53 pathway in cancer. The project also provides evidence for the possibility of using ANN approaches as data mining tools to achieve pathway-level analysis of gene expression data.

1.11. Organisation of the thesis

The project examine the possibility of using ANN as a pathway data mining tool. First by considering single cancer type. Second by considering multiple cancer types and third by performing similar analysis using existing pathway analysis tool.

The analysis carried out in three major stages:

1. ANN algorithms to model the TP53 pathway in multiple microarray datasets (considering colorectal cancer as a case study).

2. ANN approaches to perform a comparative analysis of the TP53 pathway based on the mutation status of the TP53 gene (Mutant- versus Wild-type), using RNA sequencing data from three TCGA projects (colorectal, gastric, and pancreatic cancers).

3. Performing comparative analysis using an existing tool for pathway analysis.

1.10. Major work contributions

- Identification of concordant genes associated with known TP53 pathway in colorectal cancer.
- Characterization and discovery of the key interactions and hub drivers linked to the pathway members. By modelling each member in the pathway using artificial neural network algorithms.
- Identification of distinctive and common predictors associated with the TP53 pathway based on the mutation status of the TP53 gene in three cancer types (Colorectal, Pancreatic and gastric cancers).
- Identification of unique interactions and hub drivers associated with the TP53 pathway in missense and wild type mutation status of the TP53 gene.

CHAPTER 2
NETWORK BIOLOGY METHODOLOGIES

2.1. Introduction

This chapter discusses neural network approaches, which are a form of supervised machine learning described in section 1.3.2. The supervised approach is trained to detect predictive patterns from highly complex and noisy data in biological systems. They have been discussed here separately since they were used for the analysis that was carried out in this research. The first part explains the general biology and the principle of the network models. The second part focused on the Artificial Neural network as it is the core methodology used in the project. This includes a description of the theory behind the method, the history, and the structure of the methods. It also contains a section that explains the advantages and drawbacks of the method. The third part describes how the method was adjusted to fit the research purposes.

2.2. Network biology

Network biology is a research area that recognizes biological processes as a complex set of molecular interactions. Biological networks provide a theoretical framework to model and investigate complex interactions of various entities in biological systems, offering valuable ways to understand and visualize the interactions and functions of cellular components. An ideal network model is a graphical representation of biological components, such as genes and proteins, and their interactions in a biological system. This facilitates pattern recognition and knowledge extraction from complex data (Zhang et al., 2014). It also helps predict new components' functions, which aids drug discovery. Biological network graphs usually contain a set of nodes that represent biological entities and a set of edges that represent interactions.

Biological networks include protein-protein interaction (PPI), gene regulatory (GRN), and metabolic networks. In PPI networks, proteins are represented in nodes, and the interactions between connected proteins are represented as edges. GRN represents regulation mechanisms of gene expression, in which a node presents a gene, and an edge represents a direct link between two genes, which means that the expression of one gene is directly regulated by the other (without mediation). By contrast, the metabolic network represents chemical interactions using a graph whereby each metabolite is mapped to a node, and each reaction is linked to a direct edge labelled with an enzyme (Muzio et al., 2021).

2.3. Principles behind network models

Cellular molecules are joined together to form complex networks. The majority of such molecules are identified through high-throughput technologies. The challenge has arisen of optimally assembling such components systematically into cellular networks to answer crucial biological questions about the cellular processes that govern disease initiation and progression. For instance, a pertinent question pertaining to cellular networks is how genetic abnormalities disrupt the regulatory system and contribute to cancer development. The large-scale assembly of biological components could be achieved using data-driven computational models, which can infer biological networks and provide a global understanding of the underlying mechanisms. Pe'er and Hacohen (2011) identified three principles for inferring molecular networks from data:

1. Statistical correlation network inference can be used to infer interactions between biological entities and to determine the potential influence each entity may have on another one. This uses computer power to analyse millions of hypotheses in a matter of seconds and develop a statistical score for each candidate interaction. Bayesian networks, as described by Friedman et al. (2002), are an example of network inference applying a statistical framework. Also, module network developed by Segal et al. (2003) is based on grouping the genes into modules, and assuming that genes that belong to the same module share a regulatory programme.

2. The second principle assumes that networks are not fixed but rather respond to different internal and external signals. Irish et al. (2004) showed the influence of growth factors and cytokines by identifying unique cancer network profiles that correlate with genetics and disease outcomes.

3. Differential network strategies can detect key components that alter the network functionality and model their interaction with another component in the system.

2.4. Artificial neural network

2.4.1. General overview

ANNs were defined by Jain et al. (1996) as "massively parallel computing systems consisting of an extremely large number of simple processors with many interconnections." They are a form of machine learning using algorithms learned from patterns. They can process information in a way that mimics the biological brain network. ANNs are suitable for the analysis of non-linear complex interactions and the presentation of "real-world" problems. They have the power to handle noisy information, learn from errors, and interpret previously unseen data. These characteristics make them germane to the analysis of genomic data to gain information about complex biological systems (Lancashire et al., 2005).

2.4.2. Historical background

ANNs were first described as a simple mathematical neuron model inspired by the basic functions of biological neurons, the building blocks of the brain (McCulloch & Pitts, 1943). There are billions of neurons with different types and lengths related to their location in the body. A simple neuron consists of three major functioning units: the dendrites (which receive signals from other neurons); the cell body (which contains the nucleus and the cytoplasm); and the axon (which carries the signal to the adjacent neurons through the synaptic gap) (Basheer & Hajmeer, 2000). Rosenblatt (1958) introduced the concept of the **perceptron** as a single-layer neural network with adjustable synaptic weights and external bias designed for the classification of data where patterns are extracted only from two linearly separated classes. In other words, the perceptron function properly only if there are two classes of data, and that represents a major limitation of the model.

Rosenblatt (1958) developed a learning procedure based on the **"perceptron convergence theorem"**, which proved that perceptron learning could converge after a finite number of iterations, positioning the decision surface in the form of a hyperplane between two classes. The perceptron receives inputs (X1, X2, Xm), and links them to the synaptic weights (denoted by W1, W2, Wm), and an external bias denoted by b; consequently, the net inputs are given by:

$$X_J=\sum_{i=1}^{m} w_i x_i + b \qquad \text{(Equation 2.1)}$$

The perceptron is activated only if the net inputs are greater than the bias. It produces an output class equal to +1 when the net input is positive and a -1 output class when the net input is negative.

Subsequently, the least-mean-square (LMS) algorithm was developed by Widrow and Hoff (1960), inspired by the perceptron theorem, following linear laws in the contained linear neurons and adjustable weights. The distinctive feature of this model is that it is the first adaptive filtering algorithm that uses the steepest descent method as a form of optimization. The method estimates the results by applying an adaptive filter, in which the algorithm starts by assigning random weights and then adjusts continuously in response to statistical variations in the behaviour of the investigated network. (Haykin, 2009).

2.4.3. Structure of multilayer perceptron

The multilayer perceptron (MLP) model is the most common form of neural network structure. It processes information in three or more layers, using a nonlinear activation function; thus, it overcomes the limitation of the previously described linear models. The first layer is the **input layer**, where the input data are scaled between 0 and 1 and linked to a set of randomised weights between 0 and 1. The inputs are then propagated forward to one or more **hidden layers**, and a statistical calculation is performed in each neuron by taking the sum of the values multiplied by the weight value to generate the "neuron activation". An activation function is performed to the sum to produce an output of the network in the **output layer**. Different activation functions are used for weight calculation and adjustment of neuron hidden layers. The most widely used one is sigmoidal activation function, which can map the activation of a neuron and produce a continuous output in a range between 0 and 1. The number of hidden layers is based on the complexity of the data (Lancashire et al., 2009). Figure 2.1 represents the structure of the multilayer perceptron.

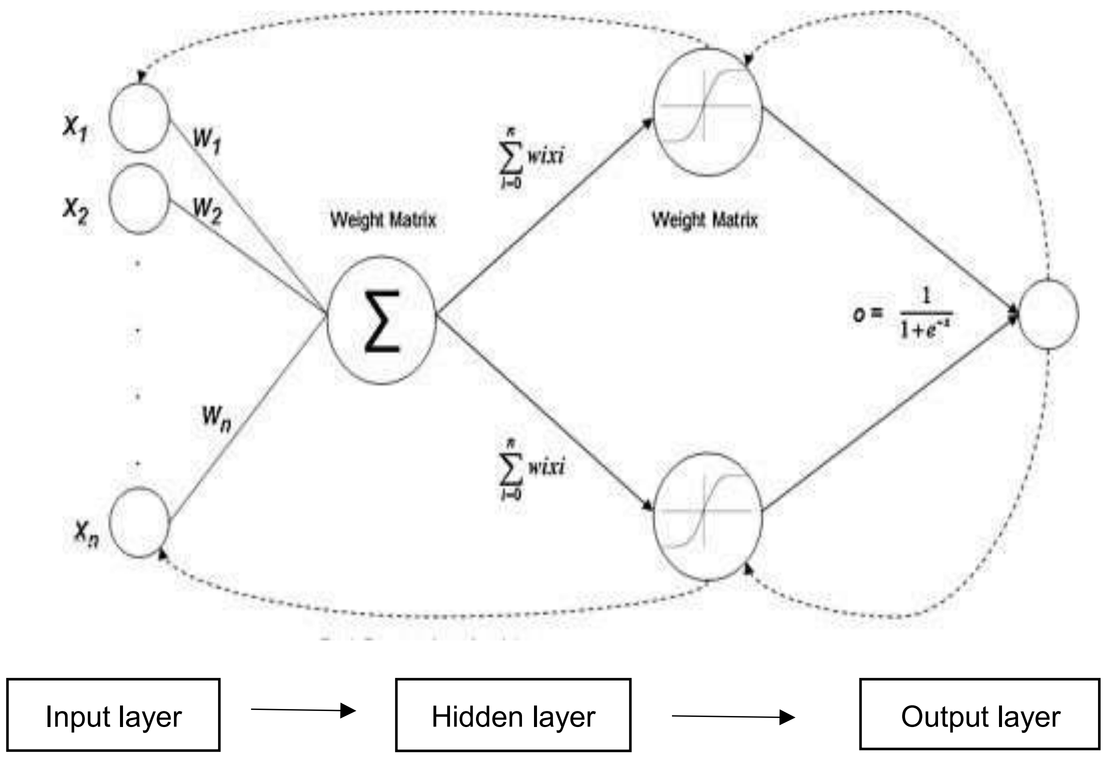


Figure 2.1: Graphic representation of the multilayer perceptron with sigmoidal activation function and backpropagation algorithm to adjust the weights

ANN with input layer, one hidden layer, and output layer

Source: adapted from Lancashire et al. (2008)

2.4.4. Learning rules

Learning is a mathematical logic that improves the performance of ANNs, enhancing their ability to perform specific tasks. Learning can be achieved through updating the internal representation of the network, which entails modification of the network structure and adjustment of the attached weights. The process of learning occurs iteratively by presenting training examples to the network. The final goal of learning is to identify the optimal set of weights that provides a learned network, which has the smallest number of errors and thus improved accuracy and capacity to define solutions closest to the expected ones. A learned network is defined by two features: (1) its ability to handle noisy, imprecise, and fuzzy information without adverse effects on response quality; and (2) its ability to generalize from the learned task to a previously unseen one.

Learning usually follows specific rules that define how the weights link the network neurons (the input to a neuron or two neurons together). The learning involves iterative adjustment of the weights and is controlled by a constant name, the learning rate. A large learning rate leads to a fast-learning process; if the learning rate is too small, the learning will be slower (Lancashire et al., 2009). There are four main types of learning rules used for ANN learning (Basheer & Hajmeer, 2000; Jain et al., 1996):

2.4.4.1. Hebbian learning rule

One of the earliest and simplest learning rules in the artificial neural network field. It is introduced in 1949 by Donal Hebb and it is often used for unsupervised learning tasks. In this rule, the weight is adjusted locally based on the activities of the neurons. It assumed that if two neighbour neurons are activated synchronously and repeatedly, then the weight between them is selectively increased. Based on Hebbian learning rule, the weight increased based on the following formula at every time step:

W_{ji} (t) = αxi (t). Yj (t) Δ

W_{ji} (t) = the rate at which the connection's weight grows at time step t.

α = constant learning rate.

x_i (t) = the pre-synaptic neuron's input value at time step t.

Yj (t) = the output of pre-synaptic neuron at same time step t.

2.4.4.2. Error-correction learning (ECL)

This is widely used in supervised learning for tasks such as classification or regression, estimating error (the difference between the predicted and actual outputs) occurrence in each training cycle, and using the correct output to adjust the connected weights, thereby leading to a gradual reduction in the overall network errors. At the beginning, the network received input data and intended output (target). The output of the network is then compared to the desired one, and the discrepancy (error) between the two is computed. After that, the network modifies its weights in a way that reduces this error, to get closer to the correct output for a particular input. Mathematically, the gradient of the error with respect to the weights, or error signal is proportional to the change in weights: $\Delta w = -\eta * \partial E/\partial w$, where η is a learning rate.

2.4.4.3. Boltzmann learning (BL)

Boltzmann learning, also known as energy-based learning or Boltzmann machines. A form of unsupervised machine learning used for various tasks including feature learning, pattern recognition and data compression. Boltzmann machines are defined by an energy function, which gives an energy value for every possible network configuration. The weights of the connections between the units and their states determine the energy of a configuration. The Boltzmann distribution relates the energy of a configuration to its probability. The goal of Boltzmann learning is to reduce the discrepancy between the model distribution that the Boltzmann machine represents and the distribution of the observed data. The learning includes adjustment of the connection weights

2.4.4.4. Competitive learning (CL)

A form of unsupervised learning suitable for tasks like clustering. In CL rule, the output neurons are compete among themselves for activation, and only one neuron is activated at any given time. Each neuron in the network represents a category or a cluster in competitive learning, where neurons compete to react to patterns in the input. The neuron (unit) whose weights are most similar to the input become active when the network receives an input pattern; other neurons remain inactive. The active neuron known as the winner, and its weights are adjusted to become more similar to the input pattern, to strengthen its capacity to respond to similar inputs in the future.

2.4.5. Backpropagation networks (BP)

As explained previously, ANN can detect and learn from errors by adjusting the connected weights. The learning occurs by using training cases to identify the best set of weights, which leads to a trained ANN model able to predict outputs closest to the expected values. For this purpose, the **backpropagation algorithm** is used, as first developed by Williams and Hinton (1986). In this method, the training of the network occurs in two phases: (1) the forward phase, in which the synaptic weights are attached to the inputs and propagate forward signals across the different layers of the network; and (2) the backward phase, in which an error is calculated by comparing the predicted output to the true output in respect to the connected weights, and the difference between the two values represent the error.

The algorithm aims to determine and minimize the error in each training cycle through the learning process. This is achieved by generating a backward signal across the different layers of the network to update the weights iteratively until no improvement in the error is observed, or a target error is reached. The error is determined as the total sum of squares based on the difference between the predicted output and the desired output, represented in the following equation:

$$\mathrm{E} = \frac{1}{2} \sum_{j=1}^{n} (\mathrm{dj} - \mathrm{yj})^{2} \qquad \text{(Equation 2.2)}$$

Where n is the number of cases, dj is the desired network output for the case j, and yj is the predicted network output for the case j. This learning process is commonly known as back-propagation (Lancashire et al., 2009). The error decreases from one training cycle (or epoch) to the next one based on the following equation:

$$\Delta(t) = \eta\delta\mathrm{k}\,\mathrm{xi} \qquad \text{(Equation 2.3)}$$

Where $\Delta w\ k\ i$ represents the weight change at the current training cycle (nth), $\delta\mathrm{k}$ represents the error in the output unit, Xj represents the weight associated with the input value, and η is a learning rate constant (which controls the size of weight change).

2.4.6. Generalization and Overfitting of ANN

The term **generalization** refers to the ability of the neural network to learn from patterns presented in the training data and to use these patterns for predicting good outputs of previously unseen cases. This occurs by using an algorithm that learns from a pattern and builds a statistical model that is able to generalize for future data (Haykin, 2009). For instance, using the amino acid sequences to predict the three-dimesional structure of proteins. Since protein structures can differ greatly even amongst proteins with similar sequences, generalization is crucial. Accurate predictions are necessary to comprehend protein function and develop treatments. The opposite of generalization is **overfitting**, which represents a major risk during the training of the neural network. Overfitting occurs when the network tries to memorize noise or unnecessary features from the training data, subsequently leading to poor generalization for similar unseen data. For example, prediction of gene expression levels in response to a variety of factors including different experimental settings, if the model picks up batch effects or noise in the training data, overfitting may result. To address the problem of overfitting, regularization techniques need to be applied during the training process (Libbrecht and Noble, 2015).

There are several regularization methods that can be chosen based on the data type or the required regularization performance. The most common one is resampling approach, where that data is split into three subsets: a **training set** used for training and optimization of the network, a **test set** for monitoring errors occurring during the training, and a **validation set** for independent validation of the trained model to produce an unbiased estimation of the network prediction performance for future cases. The training weights with the lowest error for the test subset are used for the final network model. The process of random splicing of the data into three subsets with a predetermined number of cases in each subset is known as Monte Carlo

Cross Validation (MCCV), which involves random shifting of the data between the various subset to enable confidence in prediction and reducing the risk of overfitting (Shao, 1993).

In the early stopping regularization method, the network produces signals to stop training when a predetermined set of iterations (epochs) have been completed or when the error for the validation or the test subset exceeds a certain minimum threshold. This reflects a reduction in the network performance, which is a signal of overfitting. Another simple regularization technique to address overfitting is weight decay. In this method, a penalty term is added to the error function. An example of this is to multiply the sum of the squared weights and biases by a decay constant that regulates how much the penalty should affect the resulting error. This method aims to keep weight value smaller by removing large weights (which are normally associated with over-fitted models) (Lancashire et al., 2009).

2.4.7. Optimization and selection of ANN parameters

ANN parameters require careful selection for practical applications to separate the signal from noise and to avoid signal overfitting. These include the number of hidden layers, the learning rate, and the addition of a momentum factor (used to accelerate the training process and prevent the network from being stuck in the flat region in error space or being trapped in a local minimum). A Momentum can be applied by a slight alteration to the rule of the weight update in the backpropagation algorithm by making the weight update in the current training cycle (tth) depend on the update of the previous cycle (t-1th). The momentum can be represented as follows:

$$\Delta w\,j\,i\;(n\,) = \eta\,\delta\,k\;\;x\,i\;+\alpha\,\Delta w\,j\,i\;(n\;-1) \qquad \text{(Equation 2.4)}$$

Where $x\,i$ is the input value, α is the momentum constant, η is the learning rate, Δw is the weight difference, and $\delta\,k$ the error in the output unit. The momentum constant could be between 0 ≤ α ≤ 1. The selection of the momentum value depends on the data and the problem under investigation. Thus, several experimental trials with a range of values are needed to identify the optimal value in a certain practical situation. For the microarray data analysis, a momentum of 0.5 with a learning rate of 0.1 has proven to be successful (Lancashire et al., 2005), (Lancashire et al., 2008).

The hidden layer contains hidden neurons which are required to process the weight sum based on the activation function in a forward direction. They are also needed for error propagation in a backward direction and for the update of the weights in the input layer (Haykin, 2009). The selection of the optimum size of the hidden layers is critical, as it affects network performance. Containing too many nodes leads to overfitting and poor generalization for unseen cases, while too few nodes lead to poor network performance by mistaking the non-linear inputs (Mitchell, 1997).

The common way to determine the size of hidden nodes in BP networks is through trial and error. **Constructive methods** can be used for this, whereby the network starts with a small number of nodes in the hidden layer, and new nodes are added one by one in the training face when needed. The advantage of the constructive algorithm is that the initial phase can simply set the number of hidden layers and neurons as one each. However, deciding when to add hidden neurons or connections and when to stop the addition process is difficult. Growing self-Organizing Maps (GSOM) can be considered as example of network that uses constructive method. In this approach, the network structure grow dynamically during training to adapt to the input data distribution. GSOM expands on Self Organizing Maps to enable the network to express intricate relationships in the data, new neurons are added to areas with high data densities. GSOM can be used for analysis of Omics data for clustering and visualization of gene expression profiles (Rauber and Dittenbach, 2002). **Pruning** methods start with oversized networks and subsequently iteratively eliminate irrelevant nodes during the training process (Liu et al., 2019). Feature selection-based approaches can be used to prune neural networks, by eliminating irrelevant features from the data. Pruning strategies based on feature selection can improve neural network performance in omics data analysis tasks by concentrating on the most important features, such as gene expression levels (Guyon and Elisseeff, 2003).

A network with one hidden layer and two hidden nodes can give the required optimal predictive performance for the microarray data analysis (Lancashire et al., 2009). In this project, a model with one input layer, one hidden layer with two nodes, and one output layer was used for data analysis.

2.4.8. ANN advantages and disadvantages

The table below provides a structured breakdown of the advantages and limitations of Artificial Neural Networks (ANNs)

Table 2.1: Advantages and limitations of ANNs in machine learning

ANN advantages and disadvantages	
Advantages	**Disadvantages**
Provide robust solutions for complex non-linear problems Suitable for the analysis of genomic data. Tolerate noise and handle missing and incomplete information. Ability to generalize by correctly classifying previously unseen data based on the training cases (Manning et al., 2014).	Overfitting: Network may memorize noisy data in the training set, leading to poor generalization. Time-consuming modeling of complex data: High dimensionality and complexity require more hidden layers, leading to longer training times. Inability of the algorithm to cover the global minimum: Addressed by randomizing initial weights before each training cycle (Manning et al., 2014). Lack of transparency ("black boxes"): Difficulty in understanding how certain outputs are reached based on inputs (Lancashire et al., 2008). Impact of data quality: Performance affected by high background variation and challenges of reproducibility associated with some technologies. Preprocessing helps mitigate these issues (Lancashire et al., 2009).

2.5. Stepwise ANN

ANNs have been demonstrated previously as powerful tools for data mining and pattern recognition (Bishop, 1995). However, the application of ANNs in biomedical research still needs to be improved. One of the main limitations is the ability of ANNs to cope with the high dimensionality associated with genomic data. This is known as the “curse of dimensionality,” first described by Bellman (1961) as “the exponential growth of the input space as a function of dimensionality.” For the purposes of this research, this pertains to the significant feature of a particular gene potentially being hidden among the vast number of other vectors in the data matrix. It occurs when the number of variables (genes) is higher than the number of cases (patients), which adds noise in the data space, leading to poor performance for unseen data. To overcome this issue, pre-processing and data dimensionality reduction methods have been widely applied (Bishop, 1995). However, feature extraction and generalization of such data remain challenging.

Stepwise ANN was developed in-house and was published by Professor Graham Ball and his team at Nottingham Trent University (Lancashire et al., 2005). It is capable of identifying patterns within the data in an iterative manner by finding the best single variables that perform the highest performance to classify the data regarding the question studied. This enables the building of the network by adding the following variables iteratively to improve classification performance and extract more reliable and meaningful information from complex data. For this project, Stepwise ANN approach was used for the identification of a panel of genes with the best predictive performance for a certain question by data mining the whole transcriptome. It is constructed to seek a model with the lowest predictive error by adjusting the network weights and adding the variables in an iterative manner. Moreover, ANNs have been successfully used for biomarker discovery in breast cancer (Abdel-Fatah et al., 2016). The study was done to determine variables that drive proliferation and the characteristics that go along with it in breast cancer, and to evaluate which variables related to clinical outcomes and response to therapy. ANN integrated data mining tools also used for biomarker discovery for Alzheimer disease (Dimitrios et al., 2018). In this study, ANN algorithms used to explore the difference in gene expression profiles between Alzheimer and healthy brains. Stepwise ANNs used for analysis of public data for to predict interaction between genes. Moreover, ANN`s integration approaches have been applied to genetic and MRI data to create a multimodality prediction system for personalised neoadjuvant breast cancer treatment (Abdel-Fatah et al., 2022).

The ANN architecture contains a single input layer, a single hidden layer with two hidden nodes with sigmoidal transfer function, to incorporate nonlinear function into the model whilst avoiding overfitting. Moreover, the network utilizes a feedforward backpropagation algorithm for updating the weights, and a root mean squared error value (RMS) for estimation of the prediction error. Initially, each gene from the transcriptomic data was considered as an individual input in the ANN, thus creating “n” individual models (where “n” refers to the number of genes studied in the experiment). All models are then sorted based on their RMS error for the unseen cases. The learning weights combined with the inputs are updated in the next training cycle based on the best-performing input at the previous step. Thus, the best performing input is removed for each consequent step, and the remaining n-1 inputs are used for analysis. This stepwise iterative process is implemented to achieve the most optimal predictive performance model, or until no further improvement in the model performance is observed (Lancashire et al., 2008).

For better model generalization and to improve the predictive performance, an MCCV strategy was applied. The samples were randomly divided into a training subset (for model learning), a test subset (to evaluate the performance of the model during the training), and a validation subset (for independent model testing for unseen cases), at a ratio of 60:20:20 (respectively). It has been found that 50 iterations are optimal to provide the most consistent model (with no further improvement observed using more bootstraps) (Lemetre, 2010). During this project, the setting of Stepwise ANN parameters was maintained, as shown below.

- **Stepwise ANN parameters**
- 3000 for the maximum number of epochs
- 1000 epochs window time
    - learning rate
- 0.5 momentum
- Weights initially randomized between -1 and +1
- Algorithm run for 20 independent loops for each single gene.
- Results were sorted based on the minimum average square error MSE for the test subset across the 20 loops.

2.6. Interaction algorithm

This algorithm performs an iterative calculation of the influence that multiple genes might have on a single gene. The difference between this algorithm and the Stepwise ANN is that instead of identifying genes with the best predictive performance, it predicts the influence each input gene has on the expression of a single output gene. It tests whether a given input gene can explain the expression variation of a certain targeted gene. In the beginning, one gene is selected as an output, and all the remaining genes are used as inputs, to explain the level of expression of the first gene by assigning a weighted score that is directly proportional to the intensity of each pair of genes. The process is then repeated iteratively for all genes in the expression matrix. Then the results are generated as a large matrix containing the interaction values across ten iterations.

The algorithm links each input to the output and determines the directionality of the interaction between a particular input (source) and an output (target) based on the sum of the weights in a pair-wise manner. The algorithm is known as an ANN-based inference algorithm. The structure of this algorithm previously described by Lemetre et al. (2009) contained a threelayered MLP, with one hidden layer of two nodes, one output layer, a sigmoidal transfer function used for calculation of the output, and a backpropagation algorithm to update the weights. An MCCV strategy was used prior to the training of the network in a percentage of 60:20:20 for training, test, and validation subsets. The process was repeated after re-sampling the cases (50 times). During these cycles of repeats, a correlation analysis was performed to compare the expected output values with the predicted values for all cases of the test subset. This is done by calculation of the Pearson correlation coefficient (r), which provides a level of confidence for each of the 50 repeats to predict the output. A threshold of r > 0.7 for 10 bootstraps was used to filter the most significant interactions (Lemetre et al., 2009). Used the same parameters in Stepwise analysis except for the epoch and the window timing, which decreased to 300 and 100, respectively. In this project, The ANNI and the Stepwise ANN have been utilized in combination. At the beginning, the Stepwise ANN was used to determine the genes with high predictive performance and lowest MSE for each of the investigated questions, this stage was done manually for the first set of experiments, and then the method was automated to reduce the analysis timing. The genes with the highest ranking among all investigated questions were selected to be used with ANNI. There is no ideal number of variables to choose from, but the number should increase in proportion to the question's complexity. For this project, 200 genes that were concordant between all questions were chosen for the interaction analysis. Figure 2.2 presents a schematic overview of the ANN-based data mining approaches used for analysis in this study and

Figure 2.3 presents a schematic overview of interaction algorithm used in the project

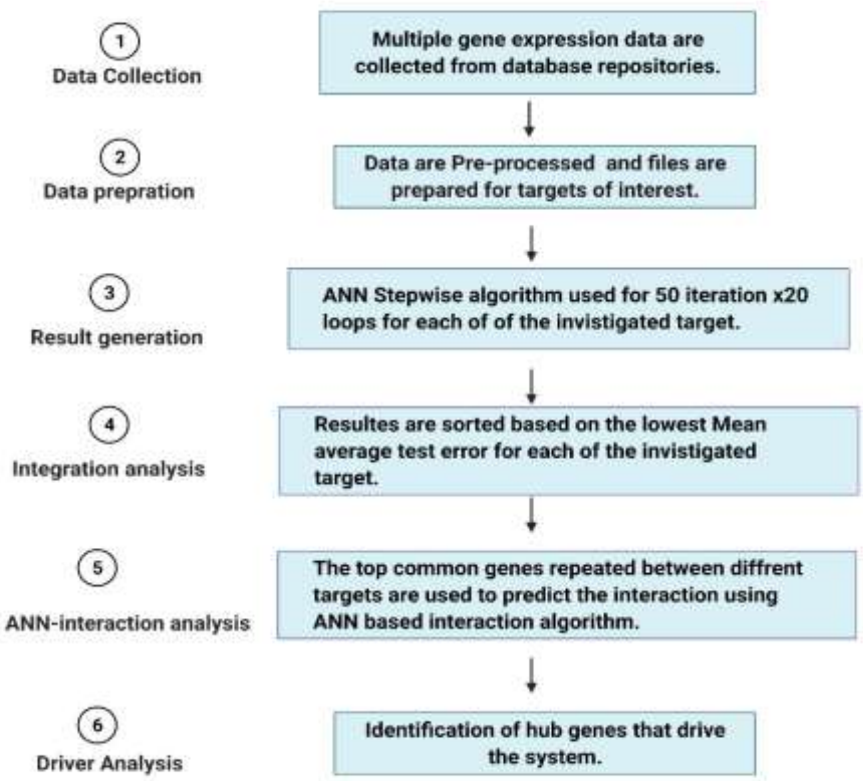

1
Data Collection
Multiple gene expression data are collected from database repositories.
2
Data prepration
Data are Pre-processed and files are prepared for targets of interest.
3
Result generation
ANN Stepwise algorithm used for 50 iteration x20 loops for each of of the invistigated target.
4
Integration analysis
Resultes are sorted based on the lowest Mean average test error for each of the invistigated target.
5
ANN-interaction analysis
The top common genes repeated between diffrent targets are used to predict the interaction using ANN based interaction algorithm.
6
Driver Analysis
Identification of hub genes that drive the system.

Figure 2.2: Schematic overview of ANN-based data mining approaches used for analysis

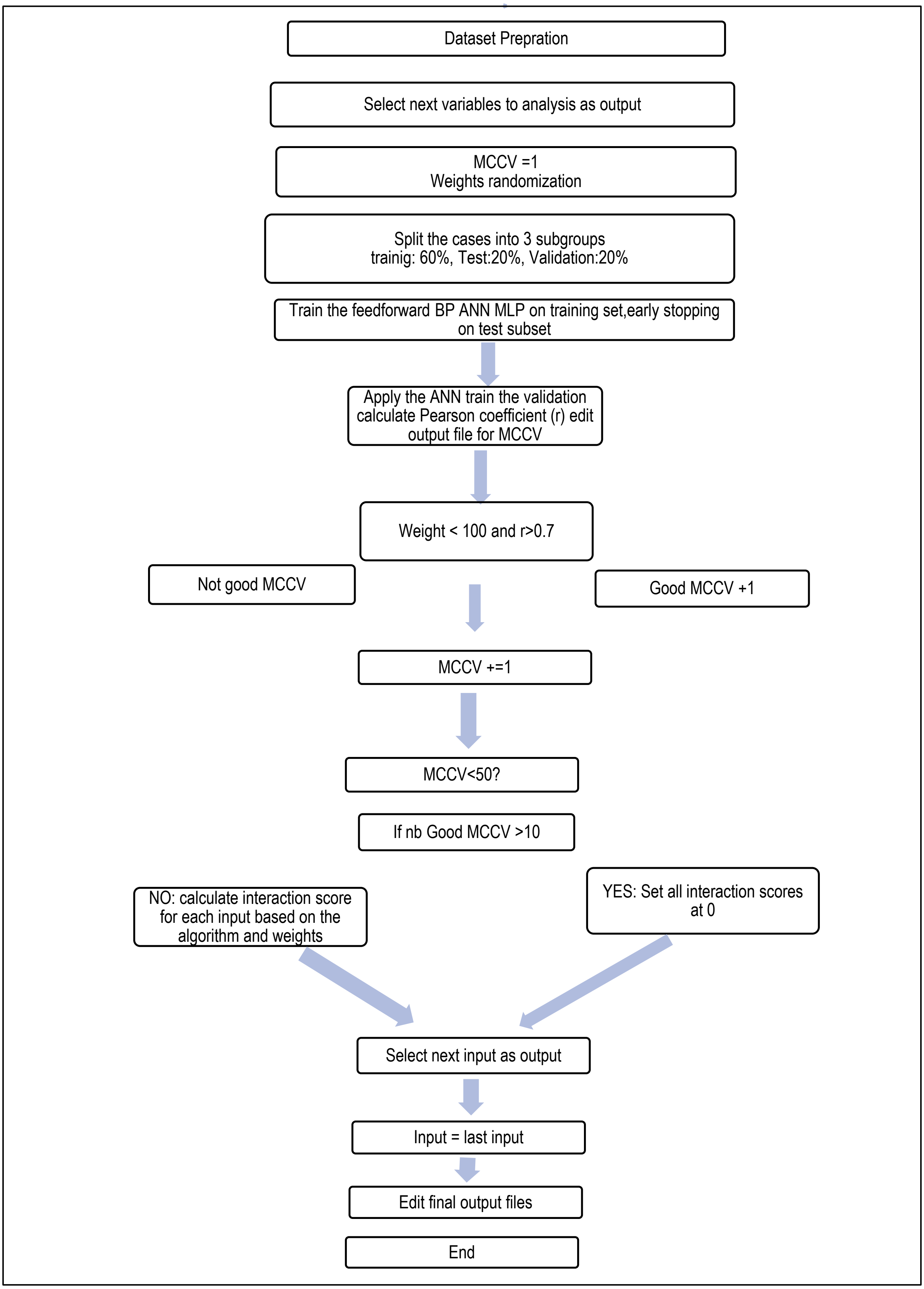

Dataset Prepration
Select next variables to analysis as output
MCCV =1
Weights randomization
Split the cases into 3 subgroups
trainig: 60%, Test:20%, Validation:20%
Train the feedforward BP ANN MLP on training set,early stopping on test subset
Apply the ANN train the validation calculate Pearson coefficient (r) edit output file for MCCV
Weight < 100 and r>0.7
Not good MCCV
Good MCCV +1
MCCV +=1
MCCV<50?
If nb Good MCCV >10
NO: calculate interaction score for each input based on the algorithm and weights
YES: Set all interaction scores at 0
Select next input as output
Input = last input
Edit final output files
End

Figure 2.3: Schematic representation of interaction algorithm used in this project

CHAPTER 3
ANN MODELLING OF TP53 PATHWAY IN COLORECTAL CANCER

3.1. Introduction

The previous chapter provides a detailed description about the Artificial Neural Network as a main approach used for the analysis carried out in this project. This chapter describes the utility of ANN-based data mining approaches for modelling the TP53 pathway in CRC. It provides a comparison between five colorectal datasets and a control dataset obtained from normal colon. The data was acquired as a metadata set from ArrayExpress database.

The first part provides a general overview of the disease, including its formation and progression. It also includes the molecular characteristic of CRC, the diagnosis and management of the disease. It also reviews the TP53 pathway in CRC. The second part concerns the study's purposes and objectives. The third is about the approaches used in the analysis. The fourth is about the essential findings and discussion. And the final part provides a summary and conclusion of the study.

3.2. Colorectal cancer (CRC)

Colorectal cancer (CRC) is the third most frequent cancer and the fourth leading cause of cancer death worldwide (Torre et al., 2016). CRC begins as an abnormal proliferation of colon epithelial tissue, known as polyps, most of which are initiated from granular cells, known as adenomas. About 10% of all adenomas are continually growing and eventually transforming into invasive adenocarcinoma. The extent of invasion of the colon wall determines the stage and prognosis of the cancer disease. Some cells metastasize to other organs via blood or lymphatic vessels (Rawal et al., 2019). There are multiple risk factors associated with CRC, including higher age-related incidence (Edwards et al., 2010), and cigarette smoking increases CRC risk. It is associated with poor prognosis (Ordonez et al., 2018).

Moreover, inflammatory bowel diseases caused mainly by ulcerative colitis (UC) increase CRC risk. A meta-analysis study by Jess et al. (2012) reported a 2.4-fold increase in the risk of CRC in patients with UC. According to a large retrospective study by Kunzmann et al. (2015), dietary fibre intake has been related to colorectal cancer incidence. They reported an inverse correlation between dietary fibre intake and the risk of adenoma formation. They pointed to cereal and fruit fibre as important nutritional constituents that reduce adenoma formation, including advanced adenoma, which is likely to progress to colorectal cancer.

However, the study found no association between dietary fibre intake and recurrent adenoma per se. It noted the potential role of genetic and lifestyle factors that make detecting association's complex in individuals with recurrent adenomas. Hereditary mutations of specific genes facilitate polyp formation and malignant transformation. Two common inherited syndromes that increase the risk of CRC are familial adenomatous polyposis (FAP) and hereditary non-polyposis colorectal cancer (HNPCC). FAP patients develop multiple adenomatous polyps. Some of which eventually progress to invasive colorectal carcinoma. Mutations in the adenomatous polyposis coli (APC) gene have long been identified as a possible cause of FAP (Groden et al., 1991; Nishisho et al., 1991). This gene functions as the "housekeeper" for cellular proliferation in colon tissue, regulating oncoprotein β-catenin. Moreover, the detection of APC mutational status is helpful for the identification of individuals at risk of developing CRC (Tsang et al., 2014). By contrast, most HNPCC occurs due to a hereditary mutation in mismatch repair genes (MMR) such as hMSH2, hMLH1, and hPMS2 (Kinzler & Vogelstein, 1996).

3.2.1. Molecular characteristics of Colorectal Cancer

CRC is a heterogeneous disease; several disease subtypes can be identified based on clinical and molecular features. Multiple genetic mutations are accumulated due to loss of genome integrity during tumour formation. Chromosomal instability (CIN), CpG island methylator phenotype (CIMP), and microsatellite instability (MSI) are common mechanisms involved in the multistep process, which develops over decades and involves several genetic events. APC mutation, followed by the sequential accumulation of other genetic mutations, including KRAS and TP53 mutations, eventually leads to CRC, as shown in Figure 3.1 (Nguyen & Duong, 2018). There are several molecular pathways involved in the formation of CRC. Disruption in the Wnt/β catenin pathway is common in CRC; it promotes cellular proliferation and inhibits differentiation. Dysfunction of the PI3K/AKT pathway is also found in CRC, and involves multiple cellular processes, including cell cycle and apoptosis. The aberrant RAS/ Raf pathway is also present in CRC, activating a series of kinase proteins responsible for transducing external signals through the plasma membrane into the nucleus (De Rosa et al., 2015)**.**

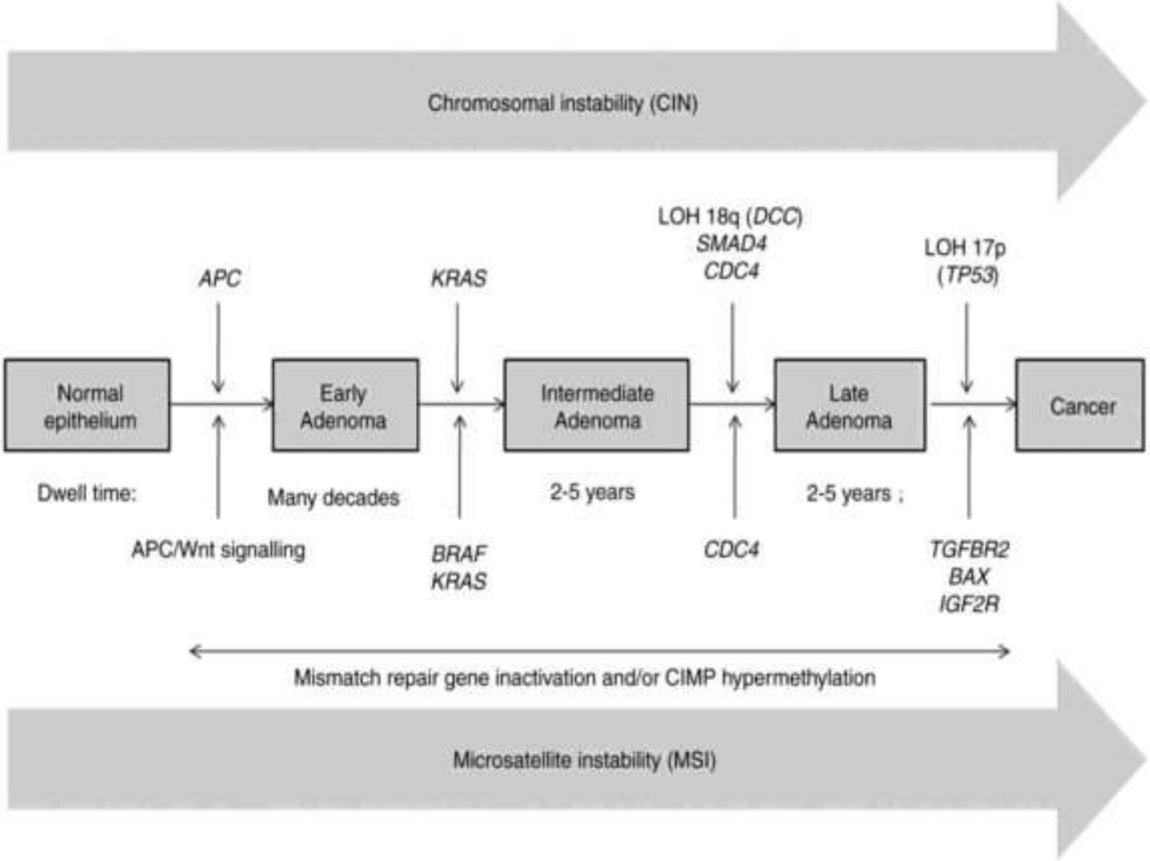


Figure 3.1: Multistep genetic model of CRC carcinogenesis

APC mutation occurs at the early adenoma stage followed by KRAS, BRAF mutations at the intermediate adenoma, then CDC4, SMAD4, LOH18Q at the late adenoma stage, and finally TP53, BAX, TGFBR2, IGF2R at the cancer stage. There are three main routes involving CIN, MSI, CIMP

Source: Adapted from Nguyen and Duong (2018). Nguyen and Duong (2018)

3.2.2. TP53 pathway in CRC

Early evidence suggested an association between the TP53 mutation and colorectal cancer. The study pointed to the TP53 mutation and the allelic loss of chromosome 17 as potential causes for the loss of tumour suppression function of the Wild-type TP53 in CRC (Baker et al., 1989). A consecutive study in Colorectal Cancer by the same group proposed the TP53 mutation as a late event that occurs during the transition from Late adenoma to carcinoma (Fearon & Vogelstein, 1990). Years later, the TP53 pathway was found to be among the most significantly altered signalling pathways in colorectal cancer tissue (compared to the normal subtype). Overexpression of the TP53 protein has been reported in 56% of colorectal cancer patients; furthermore, it has been associated with poor prognosis and reduced survival rate

(Rambau et al., 2008). A comparative study identified Mutant TP53 among the concordant genes between primary and metastatic CRC and highlighted the importance of the Mutant TP53 in tumour initiation, as well as in the progression and metastasis of CRC (Brannon et al., 2014).

Moreover, the TP53 gene has been found to be among the top-6 differentially expressed genes in grade (II and III) colon cancer, using the system biology approach. Rostami-Nejad (2019) identified a functional correlation between TP53 and other genes, including MAPK3 AKT1, and pointed out that the AKT1 gene has a negative regulation on TP53 and MAPK3 genes (i.e., overexpression of AKT1 reduces the activity of these genes). However, the results were based on one dataset, with a relatively small number of patients, emphasizing the need for more confirmation in a large-scale setting. This is one of the issues which was considered in this project.

Moreover, Slattery et al. (2019) used a statistical method based on differential expression analysis and fold change values to explore the interaction of targets from the TP53 signalling pathway in CRC compared to normal mucosa. The study supported previous reports of the influence of activated TP53 on downstream targets, and suggested that this influence may be executed via mRNA:miRNA interactions. Other genes within the TP53 pathway could also be implicated in CRC formation. Ribonucleotide reductase M2 (RRM2), a TP53 target involved in the DNA repair mechanism, has been reported as a facilitating factor for the invasion and metastasis of CRC (Lu et al., 2012). Gali-Muhtasib et al. (2008) identified a worse prognosis and increased level of Checkpoint kinase 1 (CHEK1) in advanced stages of CRC and suggested that the addition of CHEK1 inhibitors could be a promising therapeutic strategy for CRC.

3.2.3. Diagnosis and management of CRC

The standard method for diagnosis of CRC is colonoscopy followed by a histopathology examination. This method is helpful in tumour localization and polyp removal, and the technique is precise and sensitive. However, in some cases, there are difficulties regarding the preparation method and patient tolerance. Computed tomography-colonography (CT or CTC) is used as an alternative method to colonoscopy in CRC diagnosis (De Rosa et al., 2015). Moreover, the TNM tumour staging system is the most important way to the evaluation of CRC. It involves the determination of tumour invasion depth (T), Lymph node involvement (LN), and distal metastasis (M). The system is used for staging and classification of CRC into four main stages:

- Stage I and stage II – localized tumour.
- Stage III – regional spread
- Stage IV – distal spread tumour.

The main aim of the TNM staging system is to determine the severity of the disease and to guide therapy choices (Greene et al., 2008). It is also used as a prediction system for prognosis, although it has some reliability issues, especially for identifying high-risk stage II patients. Moreover, faecal occult blood tests are used for early detection of CRC in many screening programs due to their ease and low cost, but related evaluations are considered to be subjective, as various factors lead to low sensitivity rates (Alves et al., 2019).

Many therapeutic strategies are used for the management of CRC, which is mainly based on the disease stage. Complete mesocolic excision is the standard surgical procedure for managing primary CRC. In an emergency, segmental colectomy will be a better choice (De Rosa et al., 2015). Adjuvant chemotherapy is used for stage II and III CRC, and radiotherapy is used for rectal cancers. In the case of the metastatic unresectable lesion, palliative therapies to improve life quality and enhance survival could be appropriate choices.

Advances in microarray and sequencing technologies pave the way toward more personalized treatment for CRC and facilitate the discovery of prognostic and predictive biomarkers. Some of them entered the clinical practice and added value to the diagnosis and management of CRC. For example, KRAS is now used in clinical settings as a predictor for negative response to epithelial growth factor (EGFR) targeted therapy (Vacante et al., 2018).

Moreover, statistical and computational tools are assets in biomarker discovery and provide a clue about the molecular interactions in CRC. For example, Yong et al. (2018) used differential gene expression and the protein-protein network to correlate protein phosphatase two catalytic subunit alpha (PPP2CA) expression for the prognosis of CRC. The study identified the role of PPP2CA in initiating and developing CRC and its possible utility as a therapeutic target for CRC. Peterson et al. (2020) used a mathematical model to predict the incidence and timing of mutation responsible for CRC initiation. The authors assumed that the model could be used as a primary step in CRC research related to early detection and therapeutic discoveries, but it needs to be validated using human data.

While such studies add great value to the CRC research field, no previous work provides a holistic view for understanding the TP53 pathway in CRC and exploring the potential interactions between its targets. Also, recent advances in molecular biology technologies generate a vast amount of data that needs to be analysed using robust data mining tools to extract meaningful biological information. Indeed, ANN and network inference algorithms have been shown previously as valuable prediction tools, whose application can lead to significant scientific discoveries. For example, Abdel-Fatah et al. (2016) utilized ANN approaches to identify the proliferation drivers most associated with breast cancer, and identified spermassociated antigen 5 (SPAG5) to be an independent prognostic marker and a therapeutic predictor for breast cancer. The result has been used as a cornerstone for consecutive research aimed at investigating the possibility of using this biomarker in the clinical practice of breast cancer.

Following in the vein of such research directions, this study proposes ANN-based algorithms as a prediction tool for the identification of hub drivers and modelling the interactions within the TP53 pathway in CRC.

3.3. Objectives

The objectives of this chapter are to:

- Collect data using the Array-express database and identify common TP53 members using the KEGG pathway database.
- Determine the top-ranked genes related to each member within the TP53 pathway by applying Stepwise ANN algorithm in four independent CRC microarray datasets.
- Identify the concordant top-200 ranked genes between different members using Excel and R programs.
- Integrate the results and identify the common genes between different members and between different datasets.
- Perform functional and enrichment analysis for the identified genes.
- Link the identified genes to previous publications using manual literature search to identify the novel genes and previously identified ones.
- Perform network inference to predict the interaction between these common genes using ANN based interaction algorithm.

3.4. Methods

3.4.1. Data source

A total of five datasets were used for this analysis, along with sixth acting as a control, as briefly described below:

1. GSE17536 – contains data from 177 CRC patients from the H. Lee Moffitt Cancer Center in the US. It has been used to identify colon cancer patients at risk of recurrence and death.
2. GSE13294 – contains data from 155 patients diagnosed with primary colorectal adenocarcinoma at Aarhus University Hospital in Denmark. It was made available for public use in 2009.
3. GSE26682 – contains data from 300 colorectal samples from the UT MD Anderson Cancer Center in the US. The samples were processed as two batches using two different hybridization techniques. GPL570 platform data were used for consistency across all comparisons. The platform contains 175 samples collected at the time of surgical resection then the tissues were frozen for RNA isolation and microarray analysis.
4. GSE14333 – contains data from 290 primary colorectal cancers collected in two centres: (A) the Royal Melbourne Hospital in Australia (n = 162 samples); and (B) the H. Lee Moffitt Cancer Center in the US (n = 128 samples). This dataset was divided into two cohorts since the original analysis was carried out in two different centres (GSE14333-A for the Royal Melbourne Hospital; and GSE14333-B for the L-Moffitt Cancer Centre).
5. GSE4183 – a control dataset, containing data from 53 normal colon, adenoma and inflammatory bowel disease samples. This was included in the analysis as a control group.

All datasets used for this analysis were developed using Illumina Affymetrix Human Genome U133 Plus 2.0 Array. The log normalized gene expression matrices were obtained as a metadataset consisting of 20,545 transcripts, with a total number of 798 patients. ArrayExpress database was used to download the data (https://www.ebi.ac.uk/arrayexpress), under accession number E-MTAB-6698.

Table 3.1: General characteristic for the data, this table indicates the percentage of general aspects of the data including the age and the grade percentage for the CRC and the control cohorts.

	GSE17536 (n=177)	GSE13294 (n=155)	GSE26682 (n=175)	GSE14333 (290)	GSE4183 (Control cohort)		
Age, mean	65.5	65.4	65	67	Group	Number	Age
Stage I, n (%)	24(13.6)	0(0)	N/A	44(15.1)	Adenoma with high grade dysplasia	9	73.6
Stage II, n (%)	57(32.2)	46(75.4)	N/A	95(32.7)	Adenoma without dysplasia	6	65.2
Stage III, n (%)	57(32.2)	7(11.5)	N/A	93(32.0)	Inflammatory Bowel Disease	15	43.8
Stage IV, n (%)	39(22)	8(13.1)	N/A	61(21.0)	Normal colon	8	50.6

3.4.2. Stepwise ANN

In this analysis, ANN algorithm was used to identify the concordant genes related to common members of the TP53 pathway in four CRC microarray datasets. The study includes 62 targets of the TP53 pathway, identified using the KEGG pathway database (https://www.genome.jp/keg/ pathway). In the database, TP53 pathway term used in the pathway text search, then the pathway name used to identify the entry ID (map04115), then the full gene list extracted manually from the database. The complete list of the targets was added to the appendix. Each target was regarded as a separate ANN model to identify the correlated gene panel to that target. ANN algorithm was applied to identify genes with the best predictive performance for each target. Initially, the algorithm sets random weights between -1 and 1, then the weights updated continuously using a three- layered feedforward back-propagation algorithm with a momentum of 0.5, the learning rate of 0.1. Based on MCCV strategy, samples were randomized into training, test, and validation with the ratio of 60:20:20 for each subset, then the samples were re-shuffled 50 times to ensure the model generalization. Model training was run for continuous analysis for a maximum of 3,000 epochs, with a 100 epoch window time, and stopped when there is no further improvement of the root mean square value of the test subset (RMS) on a threshold of 0.01.

The process was done for a minimum of two steps over 20 independent loops for each model.

The results were then sorted and rank-ordered based on the root mean squared error (RMS) of the test subset, whereby the one with the lowest test error comes on the top of the list, and so on. The process was done for the 62 targets and for the four datasets, separately and independently. The top-200 genes for each target were extracted and merged using R programme, to identify the concordant genes across multiple repeats then across multiple targets. Figure 3.2 shows a flow diagram of stepwise ANN methodology used for the analysis.

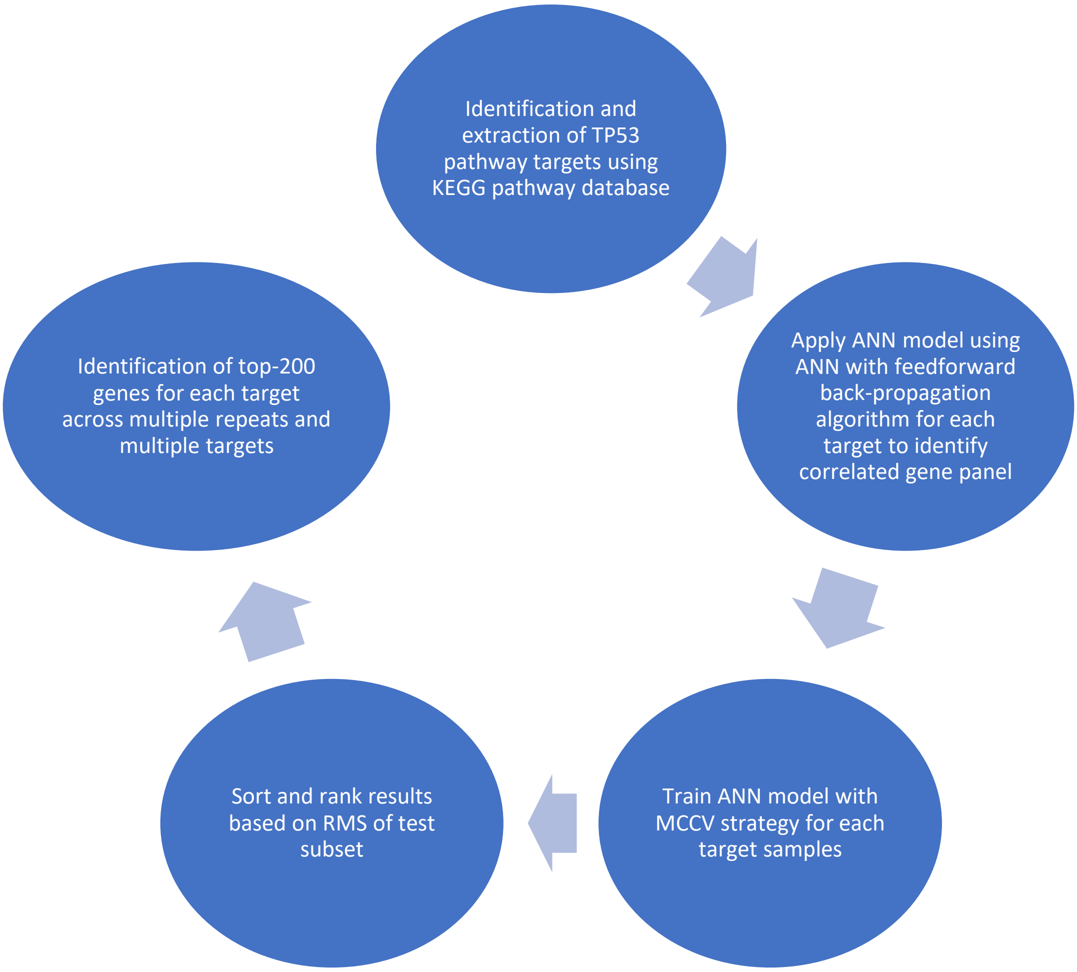

Identification and extraction of TP53 pathway targets using KEGG pathway database
Apply ANN model using ANN with feedforward back-propagation algorithm for each target to identify correlated gene panel
Train ANN model with MCCV strategy for each target samples
Sort and rank results based on RMS of test subset
Identification of top-200 genes for each target across multiple repeats and multiple targets

Figure 3.2: A flow diagram show methodology steps using Stepwise ANN for the analysis

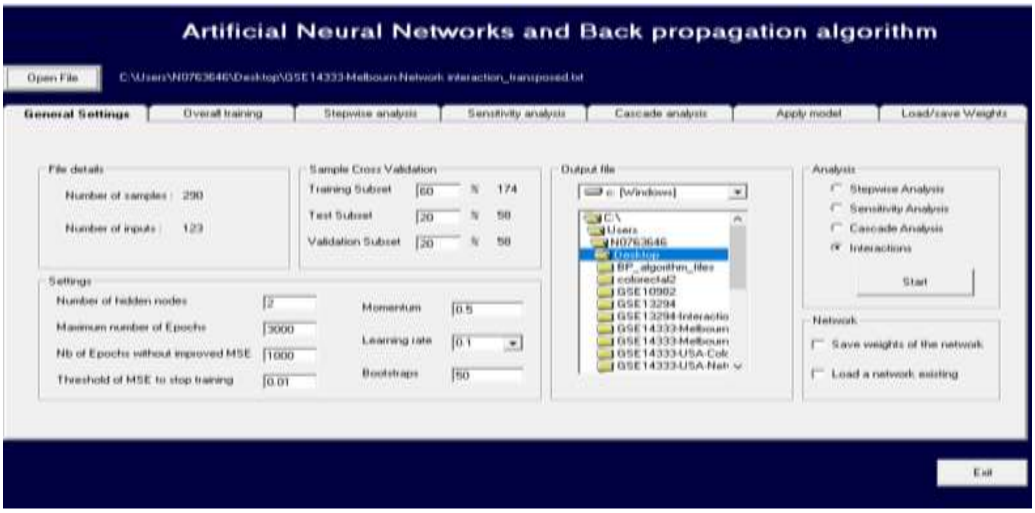
3.4.3. Network inference approach

The results of the Stepwise ANN approach for the common genes associated with the TP53 pathway were applied to the ANNI algorithm described previously in chapter 2, section 2.6. This method is used to determine the fundamental role of the genes selected by the Stepwise ANN on the TP53 pathway by quantifying genetic interaction and estimating the influence of multiple genes on a single fixed gene. In this method, each gene was considered as a single input, and the other genes (output) were used to predict the expression of that gene. The process is repeated for all genes in each dataset. The results were obtained by taking the average of 10 iterative cycles of analysis, which were then sorted to define the highest absolute values.

Genes with the highest absolute value of interactions were proposed as hub genes and therefore were considered for visualization. Cytoscape version 3.8.0, an open software programme, was used for network visualization. The programme was downloaded using (https://www.cytoscape.org/) website. The platform is used to illustrate the top-100 strong interactions for each dataset separately, whereby the genes were presented as nodes and the interaction were presented as edges. Moreover, to increase the prediction power, driver analysis was performed by calculating the sum of the weights to estimate the general influence of specific genes on the whole system. Figure 3.3 shows a visual representation of the steps used in ANN interaction approach.

The Input file consist of the original dataset that contain the expression profile of the common genes identified by the Stepwise ANN

The Input file imported to the ANN algorithm and the analysis run based on the criteria mentioned in section 2.6



Artificial Neural Networks and Back propagation algorithm

C:\Users\N0763646\Desktop\GSE14333-Melbourn-Network interaction_transposed.txt

File details
Number of samples : 290
Number of inputs : 123

Sample Cross Validation

Training Subset	60	%	174
Test Subset	20	%	58
Validation Subset	20	%	58

Settings

Number of hidden nodes	2
Maximum number of Epochs	3000
Nb of Epochs without improved MSE	1000
Threshold of MSE to stop training	0.01
Momentum	0.5
Learning rate	0.1
Bootstraps	50

Output file
c: [Windows]
C:\
Users
N0763646
Desktop
BP_algorithm_files
colorectal2
GSE10982
GSE13294
GSE13294-Interactio
GSE14333-Melbourn
GSE14333-Melbourn
GSE14333-USA-Colc
GSE14333-USA-Nat

Analysis
○ Stepwise Analysis
○ Sensitivity Analysis
○ Cascade Analysis
● Interactions

Network
□ Save weights of the network
□ Load a network existing

The output file consist of interaction matrix which generated in 10 iterative cycles of analysis calculated in a pairwise manner, where each gene was considered as a single input and the other genes were used to predict the expression of that gene.

The average and absolute value of interaction for the 10 repeats for each gene pairs calculated using average function of excel and genes with the highest absolute value visualized using Cytoscape software.

Figure 3.3: Visualization of the analysis steps used in ANN interaction approach

3.4.4. Gene ontology and enrichment analysis

To infer the biological and functional importance of the identified genes, functional enrichment analysis was done using Panther online database (http://www.pantherdb.org/). “Functional classification analysis” category was selected to compare the obtained list to a previously known gene set from literature.

3.5. Results

3.5.1. Prediction of common genes associated with known TP53 pathway members

Five datasets were used in this analysis: GSE26682, GSE13294, GSE17536, GSE14333A and GSE14333B. Each of the 62 pathway members was examined using the ANN algorithm for 20 repeats. The results were filtered using R and R studio programmes to identify candidate genes that are highly connected to each pathway member. First the data imported using (import function) of the R studio Programme, then a (data. Frame function) used to create a two dimensional data structure, then the code (data. Frame c (occurrences)) used to count the similarity in the data. Then a (write. Table function) used to create a table that link each gene with its frequency of occurrences among investigated cohorts. The results were sorted and rank-ordered based on the RMS of the test subset, the one with the lowest test error comes on the top of the list, and so on. The top-ranked 200 genes for each pathway member were identified, extracted, and then merged using cbind function in the R software program (https://www.Rproject.org/) to identify the concordant gene list of the 20 repeats. The process was done for the 62 pathway members and for the five datasets separately and independently. A final table for the top-200 ranked genes for each of the 62 members for each dataset was generated to identify concordat genes among all members. Figure 3.4 summarizes the results of the comparison for the top common genes for all datasets, indicating 33% consistency between all investigated cohorts.

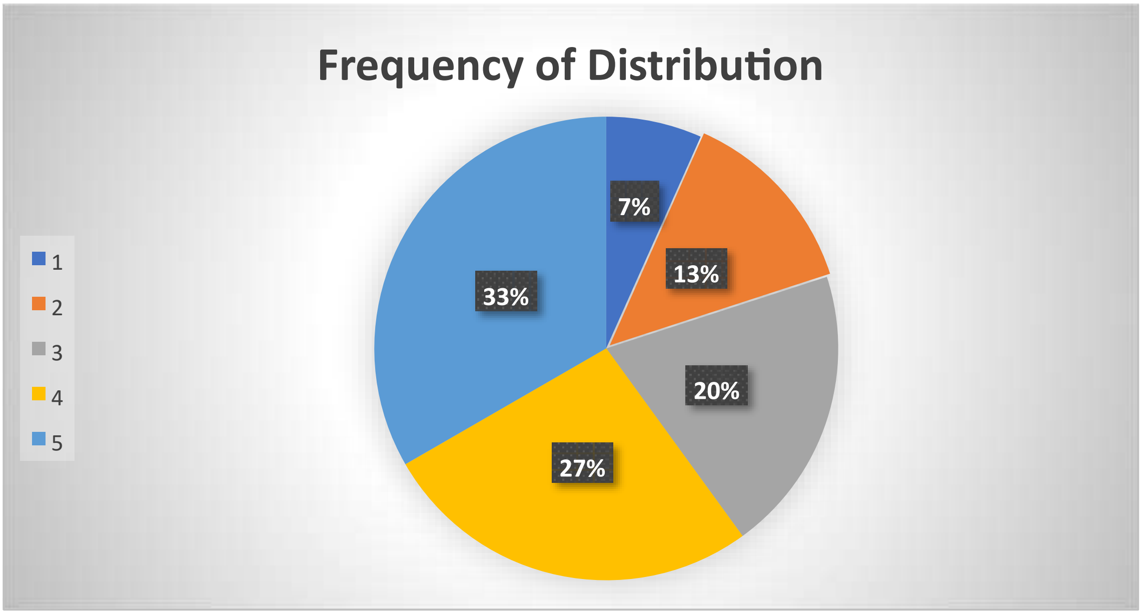


Figure 3.4: Commonality distribution frequency in the top-200 genes for the five datasets

3.5.2. Integration and ontology evaluation for common predictors

A further downstream ontological evaluation was performed to identify genes which are common between all cohorts, for a minimum of three pathway members and more. From the result of the Stepwise ANN algorithm, genes that appear common between the top-200 ranked genes for a minimum of three pathway members and for all cohorts were assumed to have more statistical power, and therefore were chosen for the ontology analysis. A minimum of 3 out of 63 used for filtration since this number was found optimum when different filtering criteria were considered. At fist a minimum of 2 was considered which result in large number of similarities that could not be handled during the next step of analysis, then a minimum of 4 and 5 members were also considered which indicate low similarities, for that a minimum of 3 members was considered optimum and used for the downstream analysis.

The results indicate the presence of **110 concordant genes** between all investigated cohorts. The probability of finding 110 common genes in the top-200 ranked genes for a minimum of 3 pathway members in 5 cohorts = $46.1255*10^{-7}$, as calculated based on the following formula:

$$(200/20545)^3*5 \qquad \text{(Equation 3.1)}$$

These common genes were then submitted to Panther online database (http://www.pantherdb.org/) for gene ontology analysis, including two categories: (1) pathway analysis, which maps the gene list to known pathways; and (2) molecular function, which indicates the events that a given protein is capable to do. The results are presented in Figures 3.5 and 3.6. It can be observed that the most significant pathways were **integrin signalling**, **inflammation mediated by chemokine**, and **apoptosis signalling pathway**. For the molecular function analysis, **extracellular matrix structural constituent**, **heparin binding**, and **glycosaminoglycan binding** were highly significant.

Moreover, the list of the common 110 genes was submitted to the EMBL-EBI European

Bioinformatics Institute database (https://www.ebi.ac.uk/) and to the National Center for Biotechnology information (https://www.ncbi.nlm.nih.gov/) to identify genes that have been previously related to the TP53 pathway in the literature. In the search box (name of a particular gene and TP53 pathway) used as a search term. The results indicated that 56 genes have been linked to the pathway in previous studies, four of which (THBS2, KIF11, CCDC68, and DDX27) were related to CRC, and eight of which were known pathway members (CCNB1, CCNB2, CDK1, CHEK1, MDM2, RRM2, DDB2, and SERPINF1). A total of 54 genes had not previously been reported (based on the search criteria, there is no paper has been published in the literature to indicate a link between a particular gene and TP53 pathway). Table 3.2 presents the gene list with their associated names and links to the original publications.

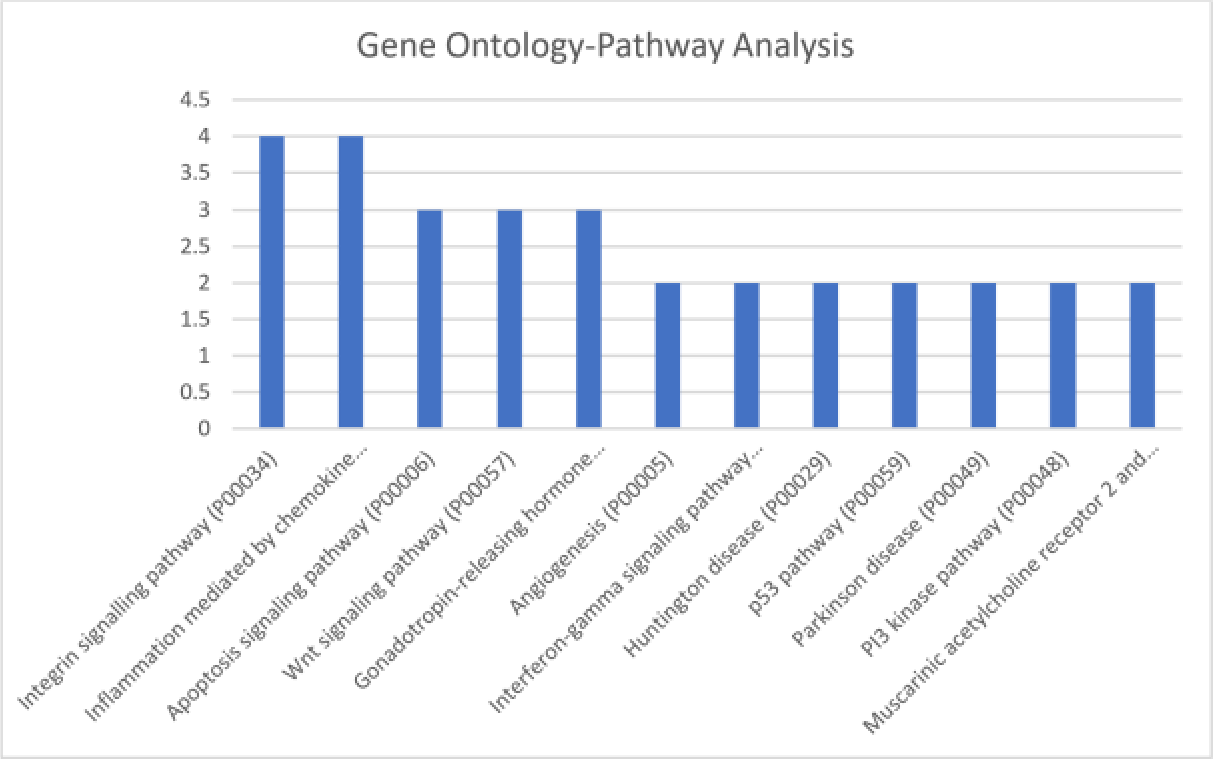


Figure 3.5: Functional classification analysis (pathway analysis) for the 110 common genes across all cohorts ranked by P-value

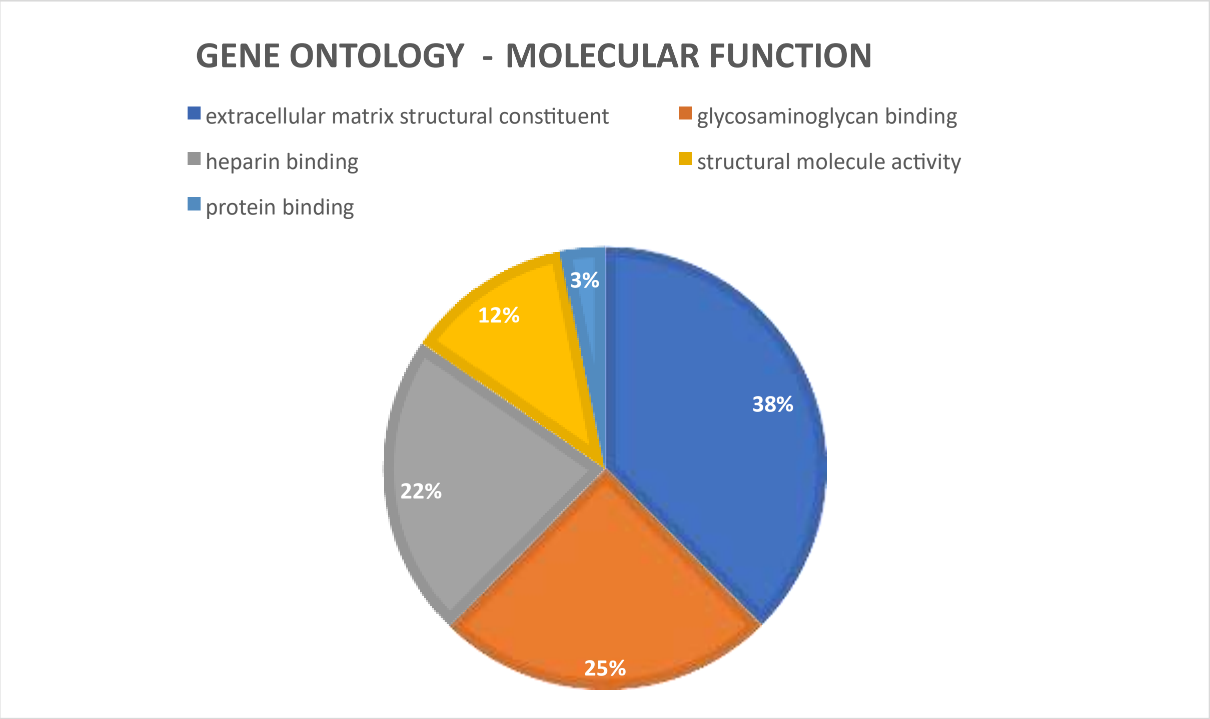


Figure 3.6: Functional classification analysis (molecular function) for the 110 common genes across all cohorts ranked by P-vale

Table 3.2: List of common genes between all investigated datasets and their correlation to the TP53 pathway in the literature (as of 23/05/2022).

Gene symbol	Gene name	No. publications	Publication
ANLN	Anilin Actin Binding Protein	2	Screening Hub Genes as Prognostic Biomarkers of Hepatocellular Carcinoma by Bioinformatics Analysis - PubMed (nih.gov)
ANXA2P2	Annexin A2 pseudogene 2	1	ANXA2P2: A Potential Immunological and Prognostic Signature in Ovarian Serous Cystadenocarcinoma via Pan-Carcinoma Synthesis. - Abstract - Europe PMC
AURKB	Aurora Kinase B	7	Evaluation of clinical value and potential mechanism of MTFR2 in lung adenocarcinoma via bioinformatics - PubMed (nih.gov)
BIRC5	Baculoviral inhibitor of apoptosis repeat-containing 5	28	Bioinformatics analysis of BIRC5 in human cancers - PubMed (nih.gov)

Table 3.2: List of common genes between all investigated datasets and their correlation to the TP53 pathway in the literature (as of 23/05/2022).

Gene symbol	Gene name	No. publications	Publication
BRCA1	Breast cancer type 1 susceptibility protein	217	Overexpression of MLF1IP promotes colorectal cancer cell proliferation through BRCA1/AKT/p27 signaling pathway. - Abstract - Europe PMC
BUB1	budding uninhibited by benzimidazoles 1	9	Cytogenetic and genetic pathways in therapy-related acute myeloid leukemia - PubMed (nih.gov)
CAND1	Cullin associated and neddylation dissociated 1	2	Partial least squares based gene expression analysis in renal failure. - Abstract - Europe PMC
CCDC68	Coiled-coil domain containing 68/ related to poor survival in colorectal cancer	1	CCDC68 predicts poor prognosis in patients with colorectal cancer: a study based on TCGA data. - Abstract - Europe PMC
BUB1B	Mitotic checkpoint serine/threonine-protein kinase BUB1 beta	7	Identification of hub genes and small molecule therapeutic drugs related to breast cancer with comprehensive bioinformatics analysis - PubMed (nih.gov)
CDC20	Cell division cycle 20	13	MDM2-P53 Signaling PathwayMediated Upregulation of CDC20 Promotes Progression of Human Diffuse Large B-Cell Lymphoma. - Abstract - Europe PMC
CCNA2	Cyclin A2	25	The p53/miRNAs/Ccna2 pathway serves as a novel regulator of cellular senescence: Complement of the canonical p53/p21 pathway - PubMed (nih.gov)
CDC6	Cell Division Cycle 6	3	p53-Dependent Regulation of Cdc6 Protein Stability Controls Cellular Proliferation - PMC (nih.gov)
CDCA5	Cell Division Cycle associated 5	1	Silencing oncogene cell division cycle associated 5 induces apoptosis and G1 phase arrest of non-small cell lung cancer cells via p53-p21 signaling pathway - PubMed (nih.gov)
DDX27	Dead Box helicase 27, related to colorectal cancer	1	DEAD-box helicase 27 plays a tumor-promoter role by regulating the stem cell-like activity of human colorectal cancer cells - PMC (nih.gov)

Table 3.2: List of common genes between all investigated datasets and their correlation to the TP53 pathway in the literature (as of 23/05/2022).

Gene symbol	Gene name	No. publications	Publication
CDKN3	Cyclin Dependent Kinase Inhibitor 3	5	YY1 suppresses proliferation and migration of pancreatic ductal adenocarcinoma by regulating the CDKN3/MdM2/P53/P21 signaling pathway - PubMed (nih.gov)
FBN1	Fibrillin 1	4	Fibrillin-1, induced by Aurora-A but inhibited by BRCA2, promotes ovarian cancer metastasis - PubMed (nih.gov)
LGALS1	Galectin 1	1	LGALS1 acts as a pro-survival molecule in AML - PubMed (nih.gov)
DNAJC9	DnaJ Heat Shock Protein Family (Hsp40) Member C9	1	Regulation of p53 and Cancer Signaling by Heat Shock Protein 40/J-Domain Protein Family Members - PMC (nih.gov)
DNMT1	DNA Methyltransferase 1	20	Human maintenance DNA (cytosine5)-methyltransferase and p53 modulate expression of p53repressed promoters - PMC (nih.gov)
PGAM1	Phosphoglycerate mutase 1	4	Phosphoglycerate Mutase 1 Activates DNA Damage Repair via Regulation of WIP1 Activity - PubMed (nih.gov)
PLK2	Polo like kinase 2	1	The p53 target Plk2 interacts with TSC proteins impacting mTOR signaling, tumor growth and chemosensitivity under hypoxic conditions - PMC (nih.gov)
DTL	Denticleless E3 Ubiquitin Protein Ligase Homolog	4	Not previously reported
RAB27B	RAB27B, member RAS oncogene family	1	Correlation Between RAB27B and p53 Expression and Overall Survival in Pancreatic Cancer - PubMed (nih.gov)
RAD51	RAD51 recombinase	60	CHK1 and RAD51 activation after DNA damage is regulated via urokinase receptor/TLR4 signaling - PubMed (nih.gov)
RPS27L	Ribosomal protein S27 like	1	Ribosomal protein S27-like and S27 interplay with p53-MDM2 axis s a target, a substrate, and a regulator - PMC (nih.gov)
ECT2	Epithelial Cell Transforming 2	11	Ect2-dependent rRNA synthesis is required for KRAS/TP53-driven lung adenocarcinoma - PMC (nih.gov)
EXOSC3	Exosome Component 3	1	Not previously reported

Table 3.2: List of common genes between all investigated datasets and their correlation to the TP53 pathway in the literature (as of 23/05/2022).

Gene symbol	Gene name	No. publications	Publication
FANC1	FA Complementation Group I	1	Not previously reported
SNRPG	Small nuclear ribonucleoprotein G	1	Downregulation of SNRPG induces cell cycle arrest and sensitizes human glioblastoma cells to temozolomide by targeting Myc through a p53-dependent signaling pathway - PubMed (nih.gov)
FDXR	Ferredoxin Reductase	3	FDXR regulates TP73 tumor suppressor via IRP2 to modulate aging and tumor suppression - PMC (nih.gov)
SPAG5	Sperm associated antigen 5	1	p53 suppression is essential for oncogenic SPAG5 upregulation in lung adenocarcinoma - PubMed (nih.gov)
KIF11	Kinesin Family Member 11, Key candidate biomarker related to Colorectal cancer.	2	Integrative analyses of molecular pathways and key candidate biomarkers associated with colorectal cancer. - Abstract - Europe PMC
ZEB2	Zinc finger E-box binding homeobox 2	2	Mutant p53-microRNA-200c-ZEB2Axis-Induced CPT1C Elevation Contributes to Metabolic Reprogramming and Tumor Progression in Basal-Like Breast Cancers. - Abstract - Europe PMC
KIF23	Kinesin Family Member 23	1	Mutation analysis and copy number alterations of KIF23 in non-small-cell lung cancer exhibiting KIF23 overexpression - PMC (nih.gov)
KIF2C	Kinesin Family Member 2C	3	Large-Scale Transcriptome Data Analysis Identifies KIF2C as a Potential Therapeutic Target Associated With Immune Infiltration in Prostate Cancer - PMC (nih.gov)
KIF4A	Kinesin Family Member 4A	1	Upregulate KIF4A Enhances Proliferation, Invasion of Hepatocellular Carcinoma and Indicates poor prognosis Across Human Cancer Types - PMC (nih.gov)
MAD2L1	Mitotic arrest deficient 2-like 1	11	Pathological significance of MAD2L1 in breast cancer: an immunohistochemical study and meta-analysis - PMC (nih.gov)

Table 3.2: List of common genes between all investigated datasets and their correlation to the TP53 pathway in the literature (as of 23/05/2022).

Gene symbol	Gene name	No. publications	Publication
MCM2	Minichromosome Maintenance Complex Component 2	10	MCM2 promotes the proliferation, migration and invasion of cholangiocarcinoma cells by reducing the p53 signaling pathway - PubMed (nih.gov)
MELK	Maternal Embryonic Leucine Zipper Kinase	2	Inhibition of MELK produces potential anti-tumour effects in bladder cancer by inducing G1/S cell cycle arrest via the ATM/CHK2/p53 pathway - PMC (nih.gov)
NCAPH	Non-SMC Condensin I Complex Subunit H	1	NCAPH plays important roles in human colon cancer - PMC (nih.gov)
ORC1	Origin Recognition Complex Subunit 1	1	Ubiquitylation, phosphorylation and Orc2 modulate the subcellular location of Orc1 and prevent it from inducing apoptosis - PMC (nih.gov)
PRC1	Protein Regulator Of Cytokinesis 1	7	Expression of the cytokinesis regulator PRC1 results in p53pathway activation in A549 cells but does not directly regulate gene expression in the nucleus - PubMed (nih.gov)
RNASEH2A	Ribonuclease H2 Subunit A	2	Prognostic Value of RNASEH2A-, CDK1-, and CD151-Related Pathway Gene Profiling for Kidney Cancers - PMC (nih.gov)
SHCBP1	SHC Binding and Spindle Associated 1	1	SHCBP1 promotes tumor cell proliferation, migration, and invasion, and is associated with poor prostate cancer prognosis - PubMed (nih.gov)
TUBB6	Tubulin Beta 6 Class V	1	Bioinformatics Analysis Discovers Microtubular Tubulin Beta 6 Class V (TUBB6) as a Potential Therapeutic Target in Glioblastoma - PMC (nih.gov)
TYMS	Thymidylate Synthetase	21	Limits to TYMS and TP53 genes as predictive determinants for fluoropyrimidine sensitivity and further evidence for an RNA-based toxicity as a major influence - PMC (nih.gov)
UBE2S	Ubiquitin Conjugating Enzyme E2 S	3	UBE2S enhances the ubiquitination of p53 and exerts oncogenic activities in hepatocellular carcinoma - PubMed (nih.gov)
UBE2T	Ubiquitin Conjugating Enzyme E2 T	3	UBE2T promotes autophagy via the p53/AMPK/mTOR signaling pathway in lung adenocarcinoma - PubMed (nih.gov)

Table 3.2: List of common genes between all investigated datasets and their correlation to the TP53 pathway in the literature (as of 23/05/2022).

Gene symbol	Gene name	No. publications	Publication
ZWINT	ZW10 Interacting Kinetochore Protein	2	Hypoxia-Induced ZWINT Mediates Pancreatic Cancer Proliferation by Interacting With p53/p21 - PMC (nih.gov)
FERMT2	FERM Domain Containing Kindlin 2	1	The Kindlin2-p53-SerpinB2 signaling axis is required for cellular senescence in breast cancer | Cell Death & Disease (nature.com)
THBS2	Thrombospondin 2, has immunological role in CRC	2	Prognostic and Immunological Role of THBS2 in Colorectal cancer - PMC (nih.gov)
EXO1	Exonuclease 1	4	Exonuclease 1 is a Potential Diagnostic and Prognostic Biomarker in Hepatocellular Carcinoma - PubMed (nih.gov)
MCMBP	Minichromosome Maintenance Complex Binding Protein	1	MCMBP promotes the assembly of the MCM2–7 hetero-hexamer to ensure robust DNA replication in human cells - PMC (nih.gov)
CKS1B	CDC28 Protein Kinase Regulatory Subunit 1B	1	CKS1B as Drug Resistance-Inducing Gene—A Potential Target to Improve Cancer Therapy - PMC (nih.gov)
MCM4	Minichromosome Maintenance Complex Component 4	1	Identification, validation, and targeting of the mutant p53-PARPMCM chromatin axis in triple negative breast cancer - PubMed (nih.gov)
E2F8	E2F Transcription Factor 8	2	The atypical E2F family member E2F7 couples the p53 and RB pathways during cellular senescence - PMC (nih.gov)
CCNB1	Cyclin B1		Known pathway member
CCNB2	Cyclin B2		Known pathway member
CDK1	Cyclin Dependent Kinase 1		Known pathway member
CHEK1	Checkpoint Kinase 1		Known pathway member
DDB2	Damage Specific DNA Binding Protein 2		Known pathway member
MDM2	MDM2 Proto-Oncogene		Known pathway member
RRM2	Ribonucleotide Reductase Regulatory Subunit M2		Known pathway member

Table 3.2: List of common genes between all investigated datasets and their correlation to the TP53 pathway in the literature (as of 23/05/2022).

Gene symbol	Gene name	No. publications	Publication
SERPINF1	Serpin Family F Member 1		Known pathway member
DCUN1D5	Defective In Cullin Neddylation 1 Domain Containing 5		No previous report indicates a relation between the gene and the TP53 pathway based on the search criteria.
DLGAP1	DLG Associated Protein 1		No previous report indicates a relation between the gene and the TP53 pathway based on the search criteria.
CDCA8	Cell Division Cycle associated 8		No previous report indicates a relation between the gene and the TP53 pathway based on the search criteria.
FEN1	Flap Structure-Specific Endonuclease 1		No previous report indicates a relation between the gene and the TP53 pathway based on the search criteria.
GINS2	GINS Complex Subunit 2		No previous report indicates a relation between the gene and the TP53 pathway based on the search criteria.
HAT1	Histone Acetyltransferase 1		No previous report indicates a relation between the gene and the TP53 pathway based on the search criteria.
KCTD9	Potassium Channel Tetramerization Domain Containing 9		No previous report indicates a relation between the gene and the TP53 pathway based on the search criteria.
KIF18A	Kinesin Family Member 18A		No previous report indicates a relation between the gene and the TP53 pathway based on the search criteria.
MCM10	Minichromosome Maintenance 10 Replication Initiation Factor		No previous report indicates a relation between the gene and the TP53 pathway based on the search criteria.
MCM3	Minichromosome Maintenance Complex Component 3		No previous report indicates a relation between the gene and the TP53 pathway based on the search criteria.
MND1	Meiotic Nuclear Divisions 1		No previous report indicates a relation between the gene and the TP53 pathway based on the search criteria.
NCAPD3	Non-SMC Condensin II Complex Subunit D3		No previous report indicates a relation between the gene and the TP53 pathway based on the search criteria.

Table 3.2: List of common genes between all investigated datasets and their correlation to the TP53 pathway in the literature (as of 23/05/2022).

Gene symbol	Gene name	No. publications	Publication
NCAPG2	Non-SMC Condensin II Complex Subunit G2		No previous report indicates a relation between the gene and the TP53 pathway based on the search criteria.
NME1	NME/NM23 Nucleoside Diphosphate Kinase 1		No previous report indicates a relation between the gene and the TP53 pathway based on the search criteria.
NUP37	Nucleoporin 37		No previous report indicates a relation between the gene and the TP53 pathway based on the search criteria.
OIP5	Opa Interacting Protein 5		No previous report indicates a relation between the gene and the TP53 pathway based on the search criteria.
PAICS	Phosphoribosylaminoimidazole Carboxylase And Phosphoribosylaminoimidazolesuccinocarboxamide Synthase		No previous report indicates a relation between the gene and the TP53 pathway based on the search criteria.
PARPBP	PARP1 Binding Protein		No previous report indicates a relation between the gene and the TP53 pathway based on the search criteria.
PBK	PDZ Binding Kinase		No previous report indicates a relation between the gene and the TP53 pathway based on the search criteria.
RACGAP1	Rac GTPase Activating Protein 1		No previous report indicates a relation between the gene and the TP53 pathway based on the search criteria.
RAD51AP1	RAD51 Associated Protein 1		No previous report indicates a relation between the gene and the TP53 pathway based on the search criteria.
RAN	RAN, Member RAS Oncogene Family		No previous report indicates a relation between the gene and the TP53 pathway based on the search criteria.
RFC2	Replication Factor C Subunit 2		No previous report indicates a relation between the gene and the TP53 pathway based on the search criteria.
RFC4	Replication Factor C Subunit 4		No previous report indicates a relation between the gene and the TP53 pathway based on the search criteria.
RFC5	Replication Factor C Subunit 5		No previous report indicates a relation between the gene and the TP53 pathway based on the search criteria.

Table 3.2: List of common genes between all investigated datasets and their correlation to the TP53 pathway in the literature (as of 23/05/2022).

Gene symbol	Gene name	No. publications	Publication
SHMT2	Serine Hydroxymethyltransferase 2		No previous report indicates a relation between the gene and the TP53 pathway based on the search criteria.
SNRPD1	Small Nuclear Ribonucleoprotein D1 Polypeptide		No previous report indicates a relation between the gene and the TP53 pathway based on the search criteria.
SNRPF	Small Nuclear Ribonucleoprotein Polypeptide F		No previous report indicates a relation between the gene and the TP53 pathway based on the search criteria.
SUV39H2	SUV39H2 Histone Lysine Methyltransferase		No previous report indicates a relation between the gene and the TP53 pathway based on the search criteria.
TIMELESS	Timeless Circadian Regulator		No previous report indicates a relation between the gene and the TP53 pathway based on the search criteria.
TIPIN	TIMELESS Interacting Protein		No previous report indicates a relation between the gene and the TP53 pathway based on the search criteria.
TK1	Thymidine Kinase 1		No previous report indicates a relation between the gene and the TP53 pathway based on the search criteria.
TOP2A	DNA Topoisomerase II Alpha		No previous report indicates a relation between the gene and the TP53 pathway based on the search criteria.
TRIP13	Thyroid Hormone Receptor Interactor 13		No previous report indicates a relation between the gene and the TP53 pathway based on the search criteria.
STON1	Stonin1		No previous report indicates a relation between the gene and the TP53 pathway based on the search criteria.
RAB31	Member RAS oncogene family		No previous report indicates a relation between the gene and the TP53 pathway based on the search criteria.
INO80C	INO80 Complex Subunit C		No previous report indicates a relation between the gene and the TP53 pathway based on the search criteria.
C18orf25	Chromosome 18 Open Reading Frame 25		No previous report indicates a relation between the gene and the TP53 pathway based on the search criteria.
PSMD9	Proteasome 26S Subunit, Non-ATPase 9		No previous report indicates a relation between the gene and the TP53 pathway based on the search criteria.

Table 3.2: List of common genes between all investigated datasets and their correlation to the TP53 pathway in the literature (as of 23/05/2022).

Gene symbol	Gene name	No. publications	Publication
SARNP	SAP Domain Containing Ribonucleoprotein		No previous report indicates a relation between the gene and the TP53 pathway based on the search criteria.
DOCK5	Dedicator Of Cytokinesis 5		No previous report indicates a relation between the gene and the TP53 pathway based on the search criteria.
HSPE1	Heat Shock Protein Family E (Hsp10) Member 1		No previous report indicates a relation between the gene and the TP53 pathway based on the search criteria.
PTGES3	Prostaglandin E Synthase 3		No previous report indicates a relation between the gene and the TP53 pathway based on the search criteria.
PPIL1	Peptidylprolyl Isomerase Like 1		No previous report indicates a relation between the gene and the TP53 pathway based on the search criteria.
TMPO	Thymopoietin		No previous report indicates a relation between the gene and the TP53 pathway based on the search criteria.
IER3IP1	Immediate Early Response 3 Interacting Protein 1		No previous report indicates a relation between the gene and the TP53 pathway based on the search criteria.

3.5.3. Network and driver analysis for common predictors

A further analysis was done using the ANNI approach described previously, to infer the interaction between the common 110 predicted genes identified in the previous stage. The analysis was performed for the list of the common 110 genes. The analysis of each set produces a matrix of ((110x (109-1)),12,210) predicted interactions. To reduce the risk of false positivises, the results were filtered by taking the average of the 10 repeats for each of the investigated sets.

To increase the prediction power and to evaluate the general influence of the predicted genes, a driver analysis was performed using the results obtained by the ANNI algorithm. This analysis was performed by taking the sum of the interaction average. The method estimates the general influence of each gene on the whole network and elucidates genes with great influence on the pathway. A parallel analysis for the top-20 strongest drivers for all datasets was performed to identify the concordant drivers and to further reduce the risk of false discovery. The results showed that **PBK**, **KIF18A**, and **ORC1** were among the top sources (influencer drivers).

Interestingly, the drivers with a high ranking in colorectal datasets tended to have a lower ranking in the control set, which gives an indication of the importance of these genes as drivers of colorectal cancer.

Moreover, **RPS27L** and **STON1** appeared among the top-ranked targets (influenced by drivers) in the diseased sets compared to the control set. There are two well-known pathway members appeared among the strongest drivers in the diseased sets compared to the control set: **MDM2** among the top sources, and **CDK1** among the top targets. The pathway member **SERPINF1** appeared as a strong source *and* target, which may reflect the importance of this gene in health and disease. **DDX27** appeared as a strong source with high rank in disease compared to normal set, it also appeared among strong targets however it got high rank in disease and normal state. The results are illustrated in Tables 3.3 and 3.4.

Table 3.3: Combined analysis for the top-20 strongest influencer sources for all cohorts

Influencer (Sources)										
Colorectal cancer Cohorts							Control Cohort			
Gene Symbol	Gene Name	Dataset	Rank	Sum of average	ABS		Gene Symbol	Rank	Sum of average	ABS
PBK	PDZ Binding Kinase	GSE14333-B	13	25.012	25.012		PBK	34	-13.059	13.05
		GSE13294	9	23.666	23.666					
		GSE26682GPL570	5	14.438	14.438					
		GSE14333-A	15	13.075	13.075					
KIF18A	Kinesin Family Member 18A	GSE13294	2	27.0008	27.0008		KIF18A	68	-8.317	8.317
		GSE14333-B	10	25.431	25.431					
		GSE14333-A	16	12.669	12.669					
		GSE26682GPL570	15	11.621	11.621					

Table 3.3: Combined analysis for the top-20 strongest influencer sources for all cohorts

Influencer (Sources)										
Colorectal cancer Cohorts							Control Cohort			
Gene Symbol	Gene Name	Dataset	Rank	Sum of average	ABS		Gene Symbol	Rank	Sum of average	ABS
DDX27	Dead Box Helicase	GSE14333-A	10	-15.586	15.586		DDX27	99	-3.032	3.032
		GSE26682GPL570	16	-11.558	11.558					
		GSE17536	13	-4.475	4.475					
		GSE13294	8	23.726	-23.726					

Table 3.3: Combined analysis for the top-20 strongest influencer sources for all cohorts. The blue colour indicates known pathway members, and the red colour indicates new candidates

Influencer (Sources)										
Colorectal cancer Cohorts							Control Cohort			
Gene Symbol	Gene Name	Dataset	Rank	Sum of average	ABS		Gene Symbol	Rank	Sum of average	ABS
ORC1	Origin Recognition Complex Subunit 1	GSE14333-B	5	28.077	28.077		ORC1	88	5.856	-5.856
		GSE13294	3	25.44	25.44					
		GSE17536	1	22.071	22.071					
		GSE14333-A	17	11.731	11.731					

Table 3.3: Combined analysis for the top-20 strongest influencer sources for all cohorts. The blue colour indicates known pathway members, and the red colour indicates new candidates

Influencer (Sources)										
	Colorectal cancer Cohorts						Control Cohort			
Gene Symbol	Gene Name	Dataset	Rank	Sum of average	ABS		Gene Symbol	Rank	Sum of average	ABS
MDM2	Mouse double minute 2 homolog	GSE14333-B	18	23.63	23.63		MDM2	35	12.696	12.69
		GSE26682GPL570	13	11.872	11.872					
SERPINF1	Serpin Family F Member 1	GSE14333-A	1	-26.668	26.668		SERPINF1	13	-18.731	18.73
		GSE26682GPL570	14	11.872	11.872					

Table 3.4: Combined analysis for the top-20 strongest influenced targets for top-20 most influenced targets

Influenced by (Targets) ***Blue colour indicates known pathway members and the red colour indicates new candidates***										
Colorectal cancer Cohorts							Control Cohort			
Gene Symbol	Gene Name	Dataset	Rank	Sum of average	ABS		Gene Symbol	Rank	Sum of average	ABS
DDX27	Dead Box Helicase	GSE14333-B	1	123.512	-123.512		DDX27	9	-61.767	61.767
		GSE26682-GPL570	1	-99.993	99.993					
		GSE13294	3	-93.487	93.487					
		GSE14333-A	10	-15.586	15.586					
		GSE17536	13	-4.475	4.475					
STON1	Stonin1	GSE14333-A	2	-103.138	103.138		STON1	58	-9.45096	9.45096
		GSE13294	1	-97.73	97.73					
		GSE14333-B	5	-90.484	90.484					
		GSE17536	17	-84.403	84.403					
		GSE26682-GPL570	10	-30.028	30.028					

Table 3.4: Combined analysis for the top-20 strongest influenced targets for top-20 most influenced targets

Influenced by (Targets) ***Blue colour indicates known pathway members and the red colour indicates new candidates***										
Colorectal cancer Cohorts							Control Cohort			
Gene Symbol	Gene Name	Dataset	Rank	Sum of average	ABS		Gene Symbol	Rank	Sum of average	ABS
RPS27L	Ribosomal Protein S27 Like	GSE26682-GPL570	11	28.762	28.762		RPS27L	82	9.576	9.576
		GSE13294	2	94.935	94.935					
		GSE14333-B	18	50.695	50.695					

Table 3.4: Combined analysis for the top-20 strongest influenced targets for top-20 most influenced targets

Influenced by (Targets) ***Blue colour indicates known pathway members and the red colour indicates new candidates***										
Colorectal cancer Cohorts							Control Cohort			
Gene Symbol	Gene Name	Dataset	Rank	Sum of average	ABS		Gene Symbol	Rank	Sum of average	ABS
		GSE14333-A	6	59.717	59.717					
		GSE17536	5	56.264	56.264					
SERPINF1	Serpin Family F Member 1	GSE26682-GPL570	15	17.126	17.126		SERPINF1	6	-64.614	64.614
SERPINF1	Serpin Family F Member 1	GSE14333-A	7	-58.57	58.57					
SERPINF1	Serpin Family F Member 1	GSE13294	19	-35.981	35.981					
SERPINF1	Serpin Family F Member 1	GSE17536	7	-7.859	7.859					
CDK1	Cyclin Dependent Kinase 1	GSE14333-A	9	55.655	55.655		CDK1	50	-20.198	20.198
		GSE14333-B	12	61.38	61.38					
		GSE17536	16	88.646	88.646					

Furthermore, the STON1 subnetwork was selected for combined analysis and visualized using Cystoscope software platform. STON1 appear as a second gene with high ranking among all investigated CRC cohorts and also has lower ranking in the control cohort compared to DDX27 which appear as a first top ranked gene in all CRC cohorts but also it appears as a top ranked in the control cohort. Another reason for choosing STON1 gene is that it has not been reported based on the literature-based search explained previously. This analysis was done by identifying and taking the average value for genes with consistent correlation with STON1 across all investigated datasets. The results of this analysis are shown in Figures 3.7 and 3.8. Figure 3.7 shows the combined disease subnetwork for STON1, which appears as a hub target that is negatively influenced by 54 genes, 4 of which are known pathway members. Figure 3.8 shows STON1 subnetwork in the normal set, with STON1 subnetwork negatively influenced by 9 genes, 4 of which are known pathway members.

Further analysis was done using the Human Protein Atlas database (The Human Protein Atlas) to identify the expression of STON1 protein in CRC. The results showed low to moderate expression of STON1 protein in 8 out of 12 investigated cases. Figure 3.9 shows digital slide images for all cases (adapted from The Human Protein Atlas).

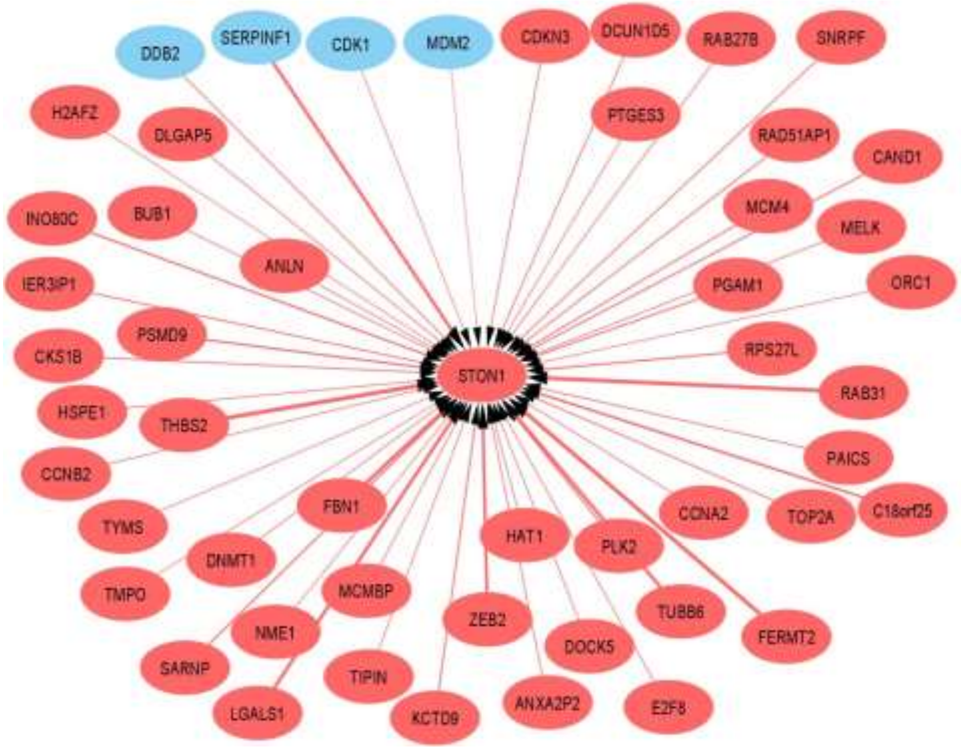


Figure 3.7: STON1 combined disease subnetwork

Shows STON1 as a hub target with 4 negative interactions with known pathway targets and 50 negative interactions with genes not related to the pathway

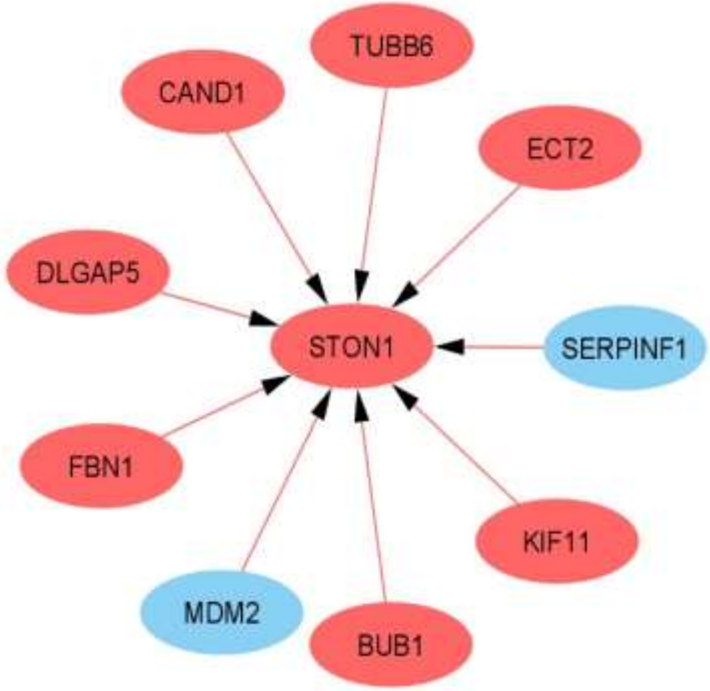
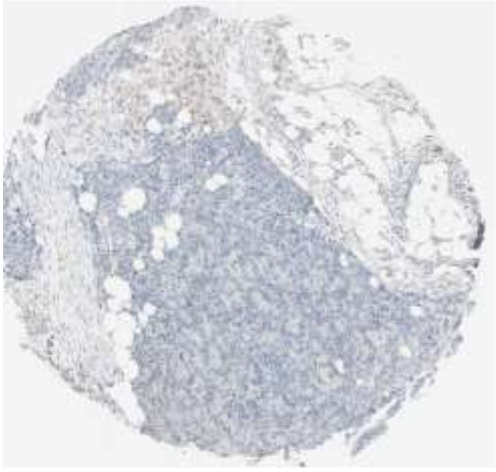
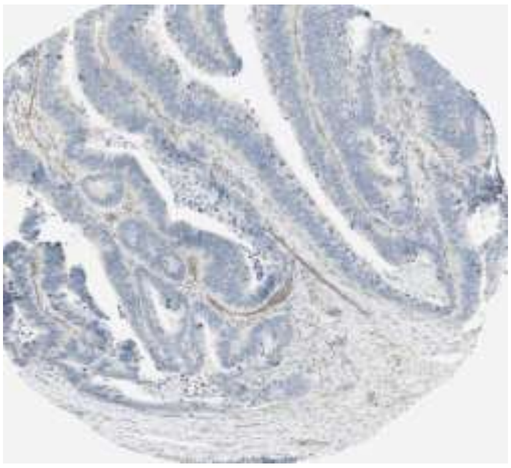


Figure 3.8: STON1 normal subnetwork

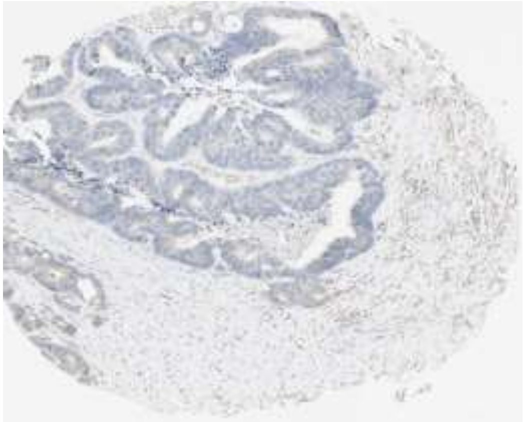
Shows two negative interactions with known pathway members and 7 negative interaction with genes not related to the pathway



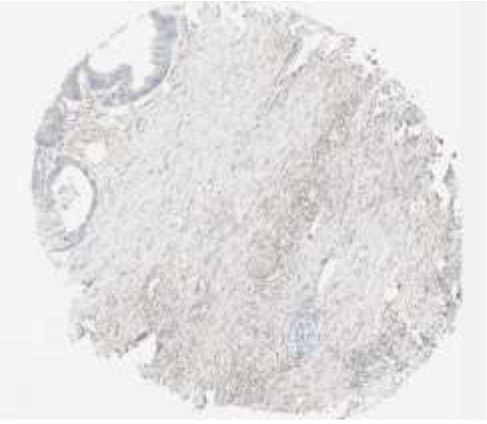
Slide 1: Negative staining



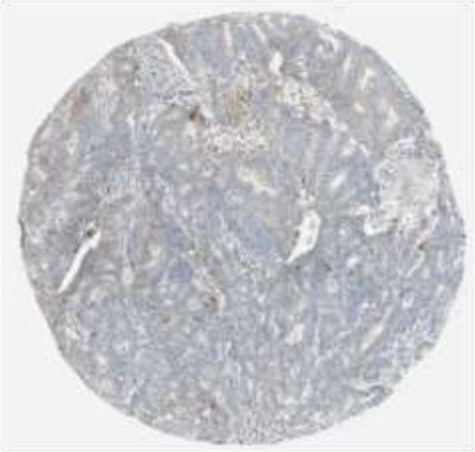
Slide 2: Negative staining



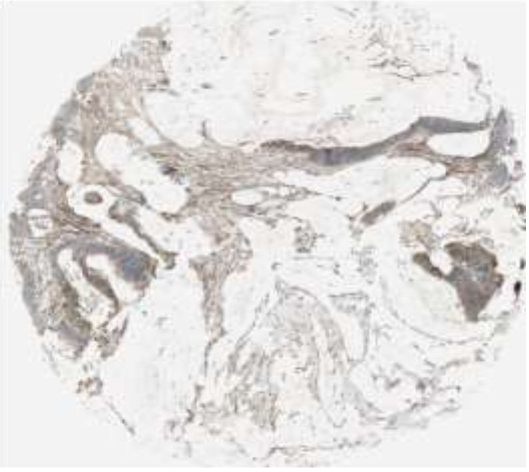
Slide 3: Negative staining



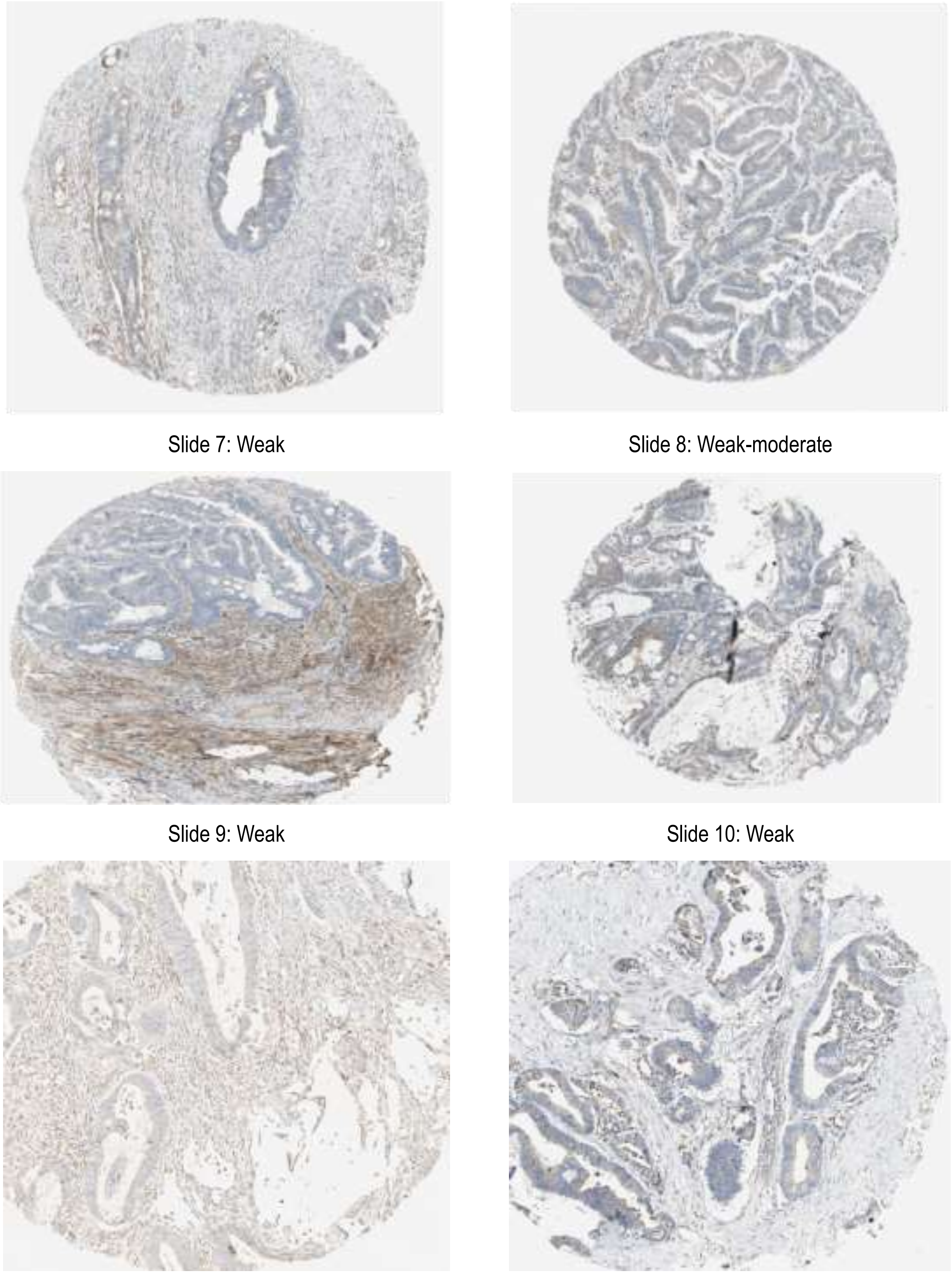
Slide 4: Negative staining



Slide 5: Weak-moderate staining



Slide 6: Moderate staining



Slide 7: Weak

Slide 8: Weak-moderate

Slide 9: Weak

Slide 10: Weak

Slide 11: Weak

Slide 12: Weak-moderate

Figure 3.9: Human Protein Atlas results for STON1 protein expression

Shows weak- moderately detected immunostaining in 8 out of 12 and negative staining in the reaming 4 examined

Adapted from Human Protein Atlas (Expression of STON1 in colorectal cancer - The Human Protein Atlas)

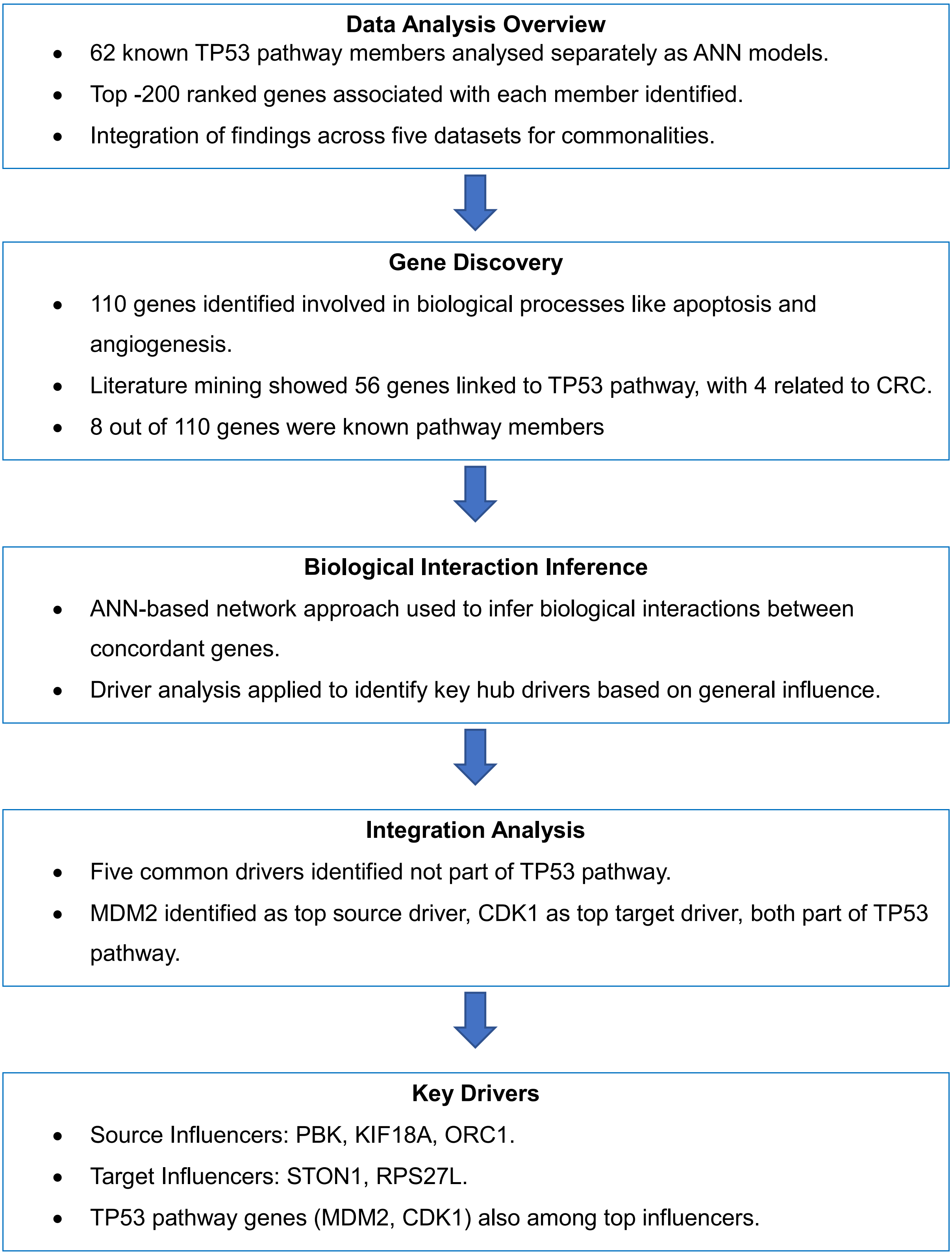
3.6. Summary and conclusion

This chapter applied ANN-based data mining to analyse four independent microarray datasets for colorectal cancer, and a control set used for comparison. The aim was to identify hub drivers and model the interactions between common members of the TP53 pathway in CRC. A random MCCV strategy was used for each model to increase the statistical power and minimize false discovery. The analysis includes 62 known pathway members, each of which was considered as a separate ANN model, in order to identify the top-200 ranked genes associated with each member. The findings were integrated to identify the commonalities across the five datasets for a minimum of three pathway members and more compared to the normal set.

The results produced a list of 110 concordant genes involved in several biological processes, including apoptosis and angiogenesis signalling pathways. A literature mining search suggested that 56 out of 110 genes linked to the TP53 pathway in this research were reported in previous studies, four of which (THBS2, KIF11, CCDC68, and DDX27) were related to CRC, while 8 out of 110 were known pathway members (CCNB1, CCNB2, CDK1, CHEK1, MDM2, RRM2, DDB2, and SERPINF1). It was found 45 out to 110 genes were not previously reported.

ANN-based network approach was used to infer the potential biological interactions between the concordant genes, and a driver analysis was used to identify the key hub drivers of the system based on their general influence. A combined analysis was then applied for all sets to find the concordant strongest drivers. The results showed that a total of five common drivers do not belong to the TP53 pathway: three of them appear as the strongest source drivers (PBK, KIF18A, and ORC1), and two of them appear as the strongest target drivers (STON1 and RPS27L). Genes that belong to the TP53 pathway also appear among the strongest drivers; MDM2 was found to be among the top strongest sources, and CDK1 was found to be among the top strongest targets. A summary of chapter overall findings provided in Figure 3.10.

This chapter provides evidence for the possibility of using ANN-based approaches as a pathway-mining tool to add knowledge and gain new insights. However, some issues related to the data, including the collection processes and lack of important information, limit the possibility of performing a deeper analysis. Chapter 4 seeks to identify the hub drivers associated with the TP53 pathway based on the mutation status of the TP53 gene.



Data Analysis Overview

- 62 known TP53 pathway members analysed separately as ANN models.
- Top -200 ranked genes associated with each member identified.
- Integration of findings across five datasets for commonalities.

Gene Discovery

- 110 genes identified involved in biological processes like apoptosis and angiogenesis.
- Literature mining showed 56 genes linked to TP53 pathway, with 4 related to CRC.
- 8 out of 110 genes were known pathway members

Biological Interaction Inference

- ANN-based network approach used to infer biological interactions between concordant genes.
- Driver analysis applied to identify key hub drivers based on general influence.

Integration Analysis

- Five common drivers identified not part of TP53 pathway.
- MDM2 identified as top source driver, CDK1 as top target driver, both part of TP53 pathway.

Key Drivers

- Source Influencers: PBK, KIF18A, ORC1.
- Target Influencers: STON1, RPS27L.
- TP53 pathway genes (MDM2, CDK1) also among top influencers.

Figure 3.10: A schematic summarise the overall findings of the chapter

CHAPTER 4
ANN DATA MINING ANALYSIS OF THE TP53 PATHWAY BASED ON TP53 MUTATION STATUS

4.1. Introduction

The previous chapter described the utility of ANN-based data mining approaches for modelling the TP53 pathway in CRC. The results indicated fundamental interactions and potential molecular drivers associated with the pathway. This chapter presents the analysis of the TP53 path based on the mutation status of the TP53 gene in three cancer types (colorectal, gastric, and pancreatic) using data from The Cancer Transcriptome Atlas. The TP53 mutations, as mentioned in section 1.5.2, represent a fundamental change during tumorigenesis, with the loss of the Wild-type TP53 functionality and the gain of additional oncogenic roles that support the survival and growth of tumour cells. The majority of TP53 alterations are Missense mutations, which attenuate the normal function of the TP53 protein. The MutantTP53 protein has a different ability to change the tumour cell proteome and transcriptome by developing new interactions with other cellular proteins, transcription regulators, and enzymes (Mantovani et al., 2019).

Several systematic projects have been dedicated to investigating Mutant TP53 in cancer, including the TCGA, which identified common alterations in the TP53 pathway, involving the

TP53 gene. Kandoth et al.'s (2013) conducted TCGA pan-cancer research on mutation frequency across 12 major cancer types. The study identified TP53 as the most mutated gene, with 42% of samples harboring TP53 mutations. Those samples were from serous ovarian and endometrial carcinomas and basal subtype breast tumours. The study used clustering analysis to identify the mutation frequency of common genes associated with different cancer types and a pairwise statistical method to identify mutual exclusive and co-occurrence among the most significantly mutated genes. Survival analysis correlated the most significant mutations with clinical outcomes, but the study did not consider pathway-level analysis. The current research selected TCGA data for analysis of the TP53 pathway, since it provides a rich source of information and contains high-quality gene expression profiles and details about mutation types. It renders the precise study of the TP53 pathway in the Mutant- and Wild-type state of gene feasible.

4.2. Chapter aims

This chapter supports the general aim of the project by providing evidence for the possibility of using ANN-based data mining approaches as pathway modelling tool through investigation of the TP53 pathway among three cancer types, to identify common and differential predictors associated with the pathway in Mutant- and Wild-type TP53. It uses the differential predictors to build interaction network and to identify differential molecular drivers associated with the pathway in the Missense TP53 mutation status. Also, this study compares ANN interaction network results with the MetaCore pathway Interactome results.

4.3. Chapter objectives

- Collect RNA sequencing data using UCSC Xena browser data hubs, and identify common TP53 pathway members using the (KEGG) pathway database.
- Identify two cohorts based on the mutation status of the TP53 gene (Mutant and Wild-type TP53 groups).
- Build ANN models for each member of the pathway in each group separately and independently.
- Determine the top-200 ranked genes associated with each member in each group.
- Identify the concordant genes between different members in each cohort.
- Define distinctive genes that are significantly associated with each group.
- Build ANN of interaction (ANNI) for specific mutation type (Missense TP53) cohort.
- Predict the key interactions and the hub differential drivers associated with the Missense TP53 cohort.
- Achieve the above objectives for three cancer types (separately and independently).

4.4. Methods

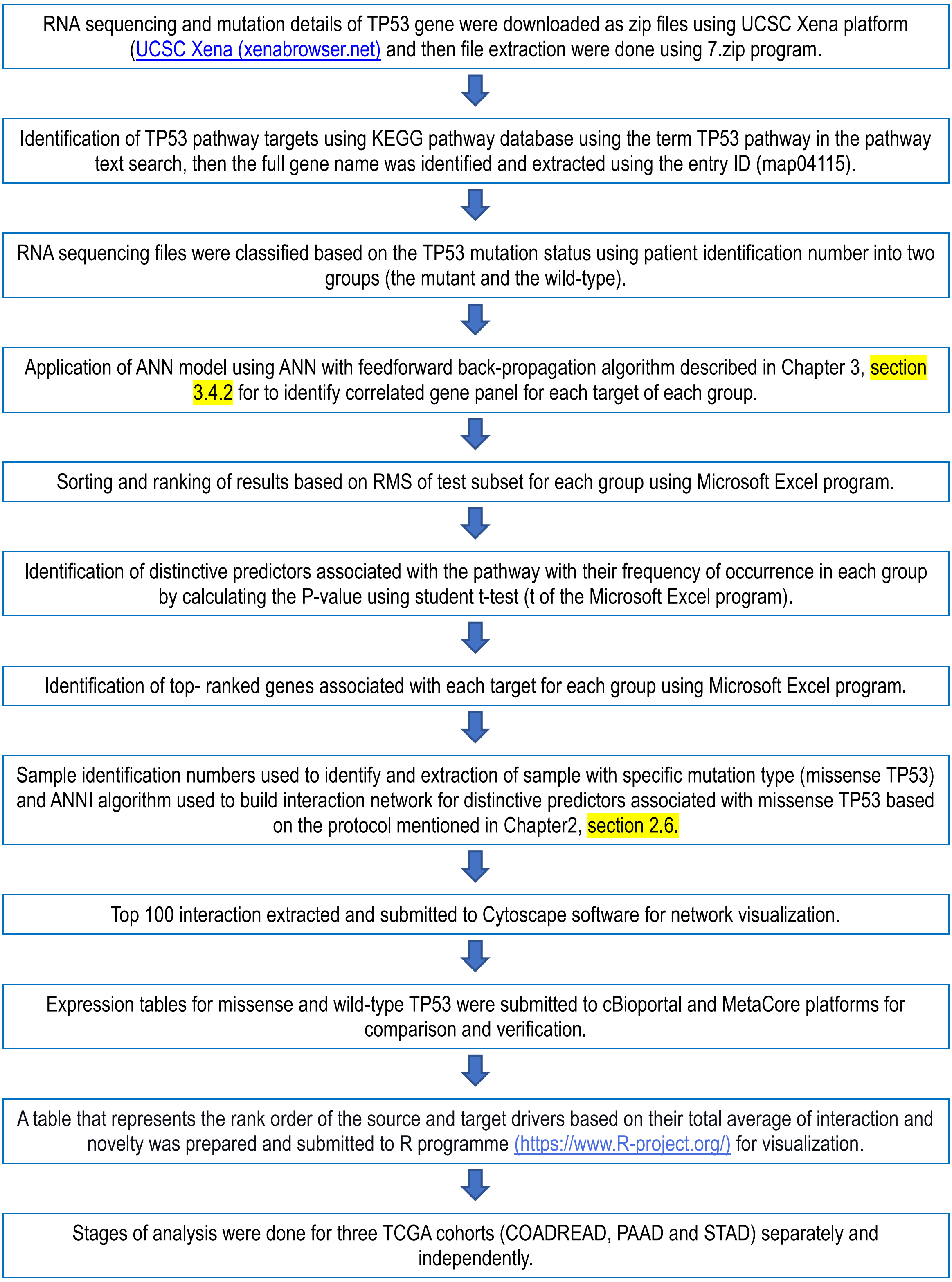
Data from three TCGA cohorts were used for this analysis for Colorectal (COADREAD), Pancreatic (PAAD), and Stomach (STAD). The data were obtained using the UCSC Xena platform (https://tcga.xenahubs.net). For the purpose of the analysis, the expression profile and the somatic mutation of the TP53 gene were included. The expression profiles for 20,531 transcripts for each cohort were downloaded as normalized data from the Xena browser gene expression RNA sequencing TCGA Hub. The gene expression was measured using Illumina HiSeq Sequencing Platform from the TCGA genome centre. The data was estimated as reads per kilobase of exon model per million mapped reads (RPKM) values, and were mapped using UCSC Xena HUGO probeMap.

- **The COADREAD cohort**: contains gene expression information for 434 colon and rectum adenocarcinoma samples from the TCGA was acquired through RNA sequencing (polyA + IlluminaHiSeq). The dataset underwent pancancer normalization, where gene expression across all TCGA cohorts were combined, averaged per gene, and then COADREAD cohort extracted.

- **The PAAD cohort:** contains 183 gene expression data for pancreatic adenocarcinoma samples from the TCGA obtained using RNA sequencing (polyA+ IlluminaHiSeq). The data has been normalized across all TCGA cohorts using a pancancer normalization method generated at UCSC, the gene expression values from RNA sequencing across all TCGA cohorts were combined, mean-centered per gene, and then the data specific to the PAAD cohort was extracted.

- **The STAD cohort:** Dataset (gene expression RNAseq - IlluminaHiSeq BC) contains RNA sequencing data for stomach adenocarcinoma of 417 samples. The data were obtained from The Cancer Genome Atlas (TCGA) and were generated using the Illumina HiSeq 2000 platform by the British Columbia Cancer Agency TCGA Genome characterization Center. Level 3 data was obtained from the TCGA Data coordination Center. This dataset provides estimates of gene expression at the transcript level, represented as RPKM.

The TP53 mutation details were obtained for each cohort from Xena browser- somatic mutation (MC3 gene-level non-silent mutation TCGA Hub). ANN and network inference approaches were used following the protocols described in the previous chapter. Figure 4.1 provide methodology details and stages of analysis.



RNA sequencing and mutation details of TP53 gene were downloaded as zip files using UCSC Xena platform (UCSC Xena (xenabrowser.net) and then file extraction were done using 7.zip program.

Identification of TP53 pathway targets using KEGG pathway database using the term TP53 pathway in the pathway text search, then the full gene name was identified and extracted using the entry ID (map04115).

RNA sequencing files were classified based on the TP53 mutation status using patient identification number into two groups (the mutant and the wild-type).

Application of ANN model using ANN with feedforward back-propagation algorithm described in Chapter 3, section 3.4.2 for to identify correlated gene panel for each target of each group.

Sorting and ranking of results based on RMS of test subset for each group using Microsoft Excel program.

Identification of distinctive predictors associated with the pathway with their frequency of occurrence in each group by calculating the P-value using student t-test (t of the Microsoft Excel program).

Identification of top- ranked genes associated with each target for each group using Microsoft Excel program.

Sample identification numbers used to identify and extraction of sample with specific mutation type (missense TP53) and ANNI algorithm used to build interaction network for distinctive predictors associated with missense TP53 based on the protocol mentioned in Chapter2, section 2.6.

Top 100 interaction extracted and submitted to Cytoscape software for network visualization.

Expression tables for missense and wild-type TP53 were submitted to cBioportal and MetaCore platforms for comparison and verification.

A table that represents the rank order of the source and target drivers based on their total average of interaction and novelty was prepared and submitted to R programme (https://www.R-project.org/) for visualization.

Stages of analysis were done for three TCGA cohorts (COADREAD, PAAD and STAD) separately and independently.

Figure 4.1: Methodology details and stages of analysis

4.5. Results and discussion

4.5.1. Analysis of TCGA-COADREAD Cohort

The Stepwise ANN approach, as explained in Chapter 2, was used in this study for the identification of common genes with the best predictive performance for each member of the pathway in the Mutant- and Wild-type TP53 cohorts, separately and independently. The results were then filtered out to identify the top-200 ranked genes that are common among all members. A comparative analysis was then applied using student t-test in the MS Excel programme, to identify common and distinctive predictors with significant differential expression between the Mutant- and Wild-type TP53 cohorts (P-value<0.05).

The analysis was first done for the TCGA-COADREAD data, then for the PADD and STAD data. The results revealed a total of 65 distinctive predictors associated with the TCGACOADREAD Mutant TP53; 79 distinctive predictors for the TCGA-COADREAD Wild-type TP53 cohort; and 37 significant concordant predictors between the two cohorts. Table 4.1 summarizes the results for the top-ranked predictors, with their frequency of appearance among all TP53 pathway members for the TCGA-COADREAD (Mutant- and Wild-type TP53 cohorts. and associated p-values.

Table 4.1: Top-ranked predictors obtained using Stepwise ANN approach

Concordant predictors for both cohorts		Distinctive predictors for TCGA-COADREAD Wild-type TP53			Distinctive predictors for TCGA-COADREAD Mutant TP53		
Gene Symbol	P-value	Gene Symbol	Frequency	P-value	Gene Symbol	Frequency	P-value
FANCB	0.0093	FDXR	8	4.77E-18	C5orf41	9	0.0035
CCT2	0.0068	C15orf23	7	0.0177	CHAF1B	8	0.0339
H2AFZ	0.0002	C3orf26	7	0.0142	ERCC6L	8	0.0040
ORC6L	0.0029	PBK	7	2.04E-10	KIAA0101	8	0.0014
AURKA	1.46E-06	RPS27L	7	5.02E-27	SERINC1	7	0.0182

Table 4.1: Top-ranked predictors obtained using Stepwise ANN approach

Concordant predictors for both cohorts		Distinctive predictors for TCGA-COADREAD Wild-type TP53			Distinctive predictors for TCGA-COADREAD Mutant TP53		
Gene Symbol	P-value	Gene Symbol	Frequency	P-value	Gene Symbol	Frequency	P-value
CDCA5	0.0402	RUVBL2	7	0.0007	TRAIP	7	0.0337
MND1	0.0366	TNFRSF10B	7	3.11E-10	AURKAIP1	6	0.0007
PA2G4	0.0294	TNFSF9	7	4.21E-11	BMPR2	6	0.0023
RAN	0.0062	TRIAP1	7	2.48E-10	CLCN7	6	2.075E-14
RFC4	0.0316	UBE2N	7	0.0032	KIF20A	6	0.0447
RFC5	0.0190	ARHGEF11	6	0.0321	MYBL2	6	1.089E-08
BUB1B	0.0151	BAT2	6	0.0153	NAP1L3	6	0.0384
CCNA2	0.0099	BOLA3	6	0.0450	PGAM5	6	0.0264
CDCA8	0.0352	C4orf46	6	0.0372	PGR	6	0.0180
CDKN3	0.0361	DDAH2	6	3.61E-06	PSAT1	6	0.0350
DBF4	0.0294	ERH	6	0.00240	RAB27A	6	2.732E-06
DSCC1	0.00098	GEMIN7	6	0.0059	RAD51	6	0.0172
E2F1	1.03E-07	MDM2	6	6.77E-31	RIF1	6	0.0451
FAM54A	0.0426	NR3C2	6	0.0020	SECISBP2L	6	0.0013
NUP37	0.0212	PRMT1	6	0.0006	TELO2	6	0.0256

Table 4.1: Top-ranked predictors obtained using Stepwise ANN approach

Concordant predictors for both cohorts		Distinctive predictors for TCGA-COADREAD Wild-type TP53			Distinctive predictors for TCGA-COADREAD Mutant TP53		
Gene Symbol	P-value	Gene Symbol	Frequency	P-value	Gene Symbol	Frequency	P-value
ORC1L	0.0320	PSMA3	6	0.0183	ZBTB4	6	0.0140
RCC1	0.0281	SKA1	6	0.00010	ACD	5	2.142E-05
SKA3	9.95E-07	SNX30	6	0.0161	ATP5F1	5	2.692E-05
SNRPF	0.0324	ZZEF1	6	0.00026	BOC	5	0.0033
SPC25	0.0422	AEN	5	4.06E-11	BRCA1	5	0.0089
TPX2	3.76E-07	AKAP13	5	0.0130	C13orf33	5	0.0190

Moreover, the distinctive predictors for the TCGA-COADREAD Mutant TP53 cohort identified in the previous stage were assumed to have more statistical power. They could be used to build network inference and to identify the key drivers which can be used to differentiate this cohort. ANNI approach was used for the analysis of patient samples that harbour TP53 Missense mutations in the TCGA-COADREAD Mutant TP53 cohort (since it represents the most frequent mutation among all types of mutations in this set, with a total number of 233 out of 242 Mutant samples). The analysis of this set produced a matrix of ((96x (95-1)) 8930 predicted interactions. The results were filtered, and a driver analysis was performed using a similar method (described in section 3.5). The top-100 interactions were presented using Cytoscape software.

The results show that SESN1, SIAH1, GADD45G, and CCNG1/TNFRF10B were key hubs, with mainly negative interactions (with genes that are not related to the pathway). SESN1 appears to be a major subnetwork containing five negative interactions with known TP53 pathway members, including TP53, SIAH1, CCNG1, BCL2, and APAF1, and one positive interaction with FAS. It also contains 14 main negative interactions with genes that are not related to the pathway, the most common of which are XPOT, TRAIP, and ZGPAT. It has two main positive interactions with LIMA1 and DUSP4. SESN1 is a member of the sestrins family, a recent report indicates a natural killer function of sestrins in one type of immune system cells (senescent-like CD8 T cells) (Pereira et al., 2020). By aligning this function of SESN1 with the presented subnetwork, the Missense TP53 may have an immune suppression role, which could be executed through its negative regulation on SESN1. For this SESN1 subnetwork was selected for presentation. LIMA1 was previously identified as a direct transcriptional target of TP53 and a possible therapeutic target for cancer (Ohashi et al., 2017). The presented subnetwork supports the association of LIMA1 with TP53 and indicates that this gene may exert its effect through a positive correlation of SESN1. Figure 4.2 shows the Cytoscape image for the top 100 interactions of the TCGA-COADREAD-MissenseTP53 mutation cohort. SESN1 mRNA protein expression was tested using the cBioPortal platform (cBioPortal for Cancer Genomics). The results indicate significantly higher expression in the Missense compared to the Wild-type cohort.

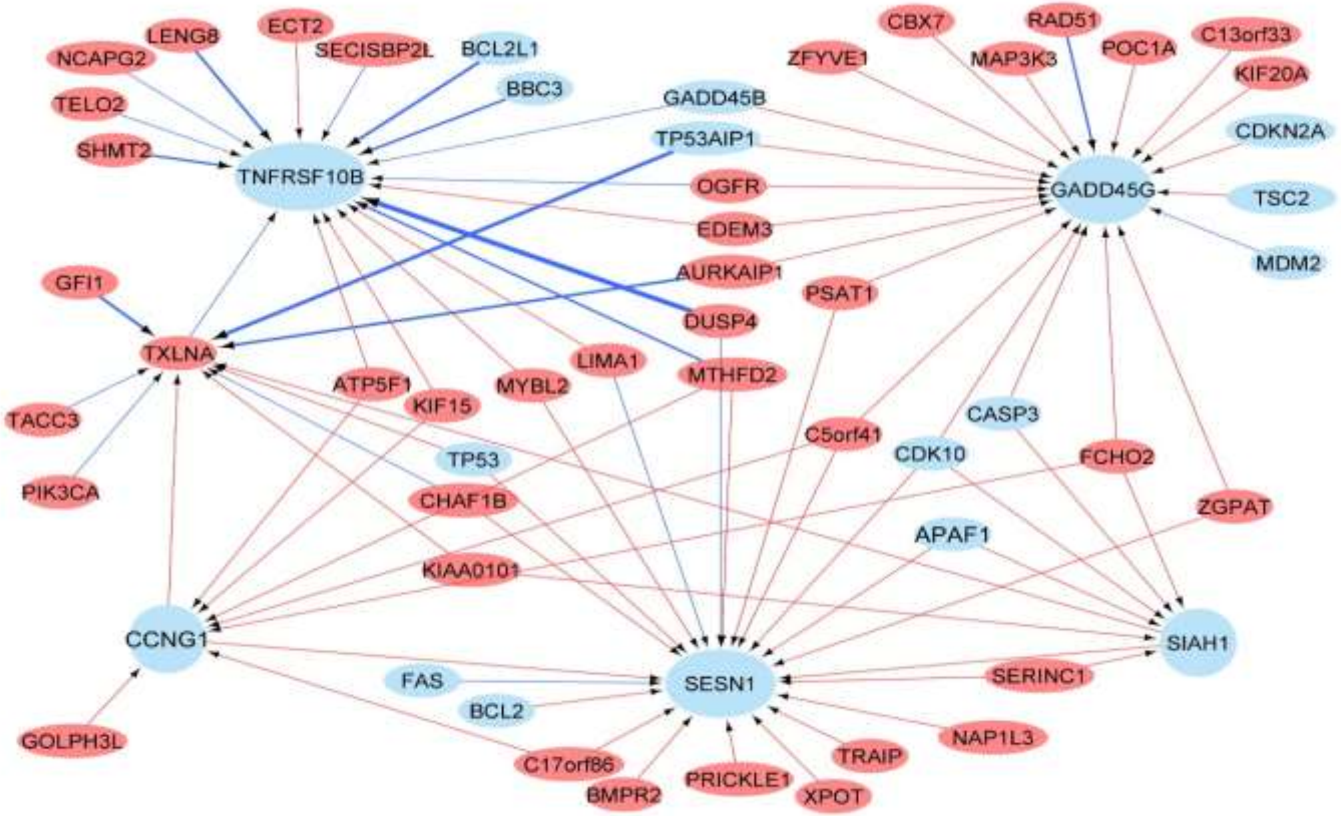


Figure 4.2: Cytoscape image for the TCGA-COADREAD-MissenseTP53 cohort

The blue nodes indicate known TP53 targets, and the red nodes indicate new candidates. The blue lines indicate positive interactions, and the red lines indicate negative interactions. Line thickness indicates interaction strength. Hub nodes are those with more than 5 interactions.

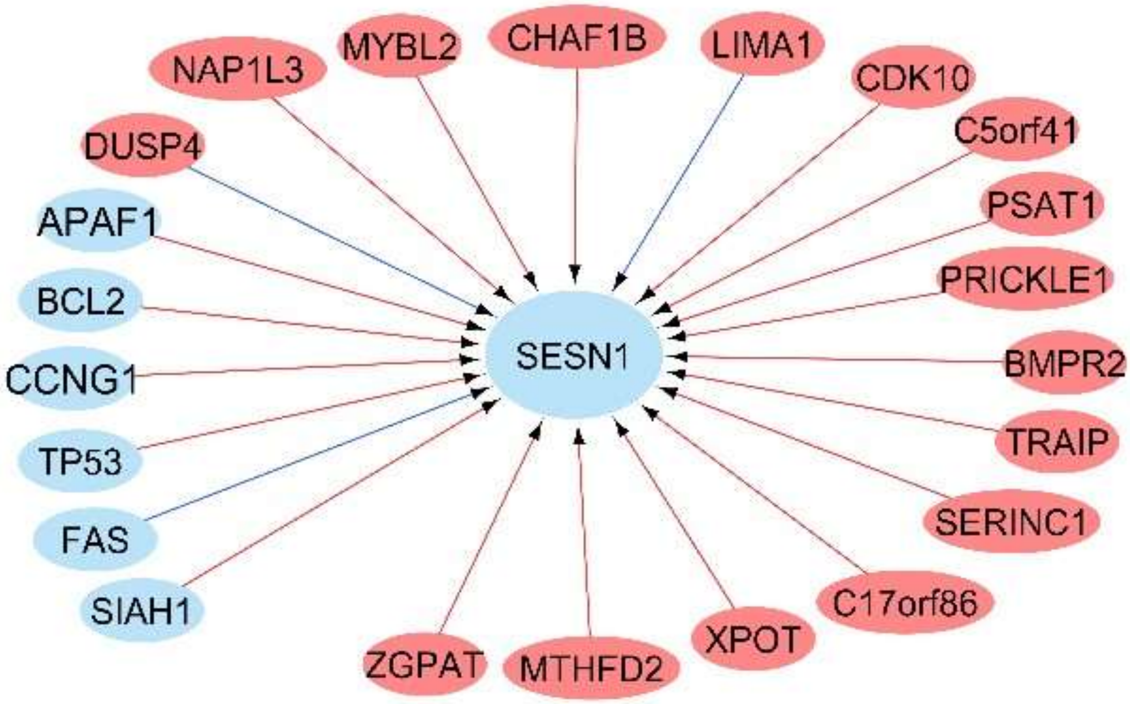


Figure 4.3: TCGA-COADREAD-MissenseTP53-SESN1 subnetwork

Contains 6 interactions with known pathway members and 16 interactions with genes not belonging to the pathway.

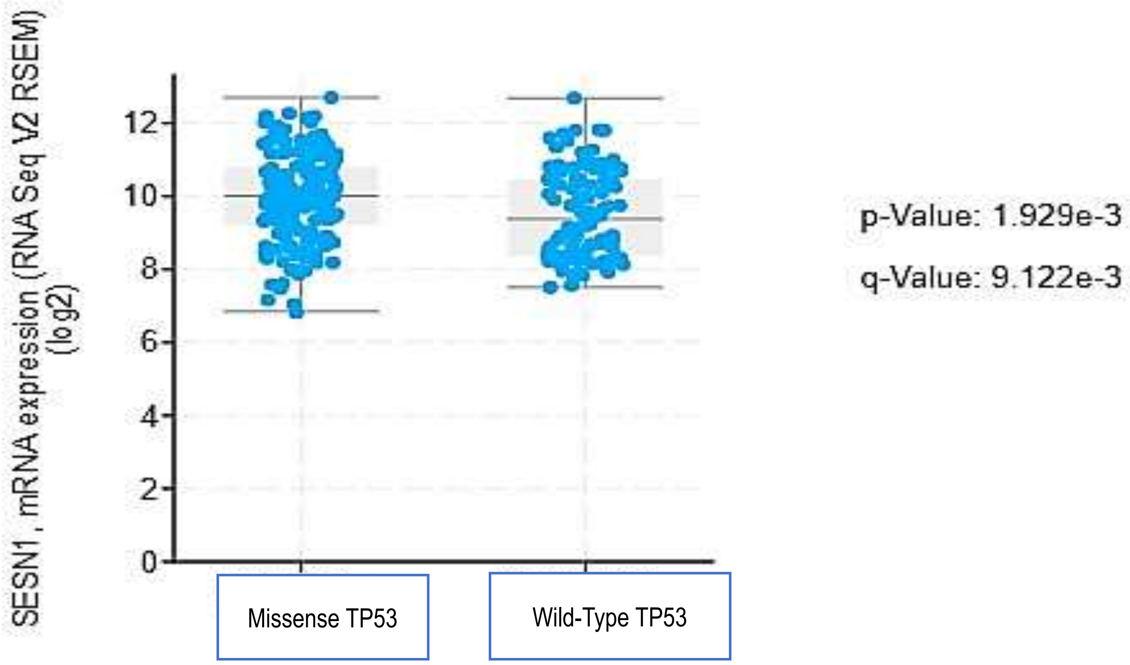


Figure 4.4: Expression of SESN1 gene in cBioPortal

The gene is more highly expressed in the TCGA-COADREAD-Missense TP53 cohort compared to the TCGA-COADREAD-WTTP53 cohort

Source: cBioPortal.

Figures 4.5 and 4.6 represent differential drivers as targets and sources for the TCGA-COADREAD-MissenseTP53 cohort. Figure 4.5 indicates the top-12 target drivers with a total positive interaction (stimulatory effect). Three of them are known TP53 pathway members (CASP3, DDB2, and APAF1) and the remaining (RIF1, EVC2, TXLNA, GFI1, NCAPD3, MAP3K3, PRC1, CENP1 and CBX7) are novel to the pathway. It also shows target drivers with a total negative interaction (inhibitory effect) representing a large pole of the interaction network (83 out of 95). A similar analysis was done for the TCGA-COADREAD-MissenseTP53 source drivers, which are represented in Figure 4.6, showing six differential source drivers with the strongest positive influence: CENP1, DUSP4, PHLDA1, KIF20A, KIF11, and TOP2A. The remaining sources have a generally negative impact on the pathway.

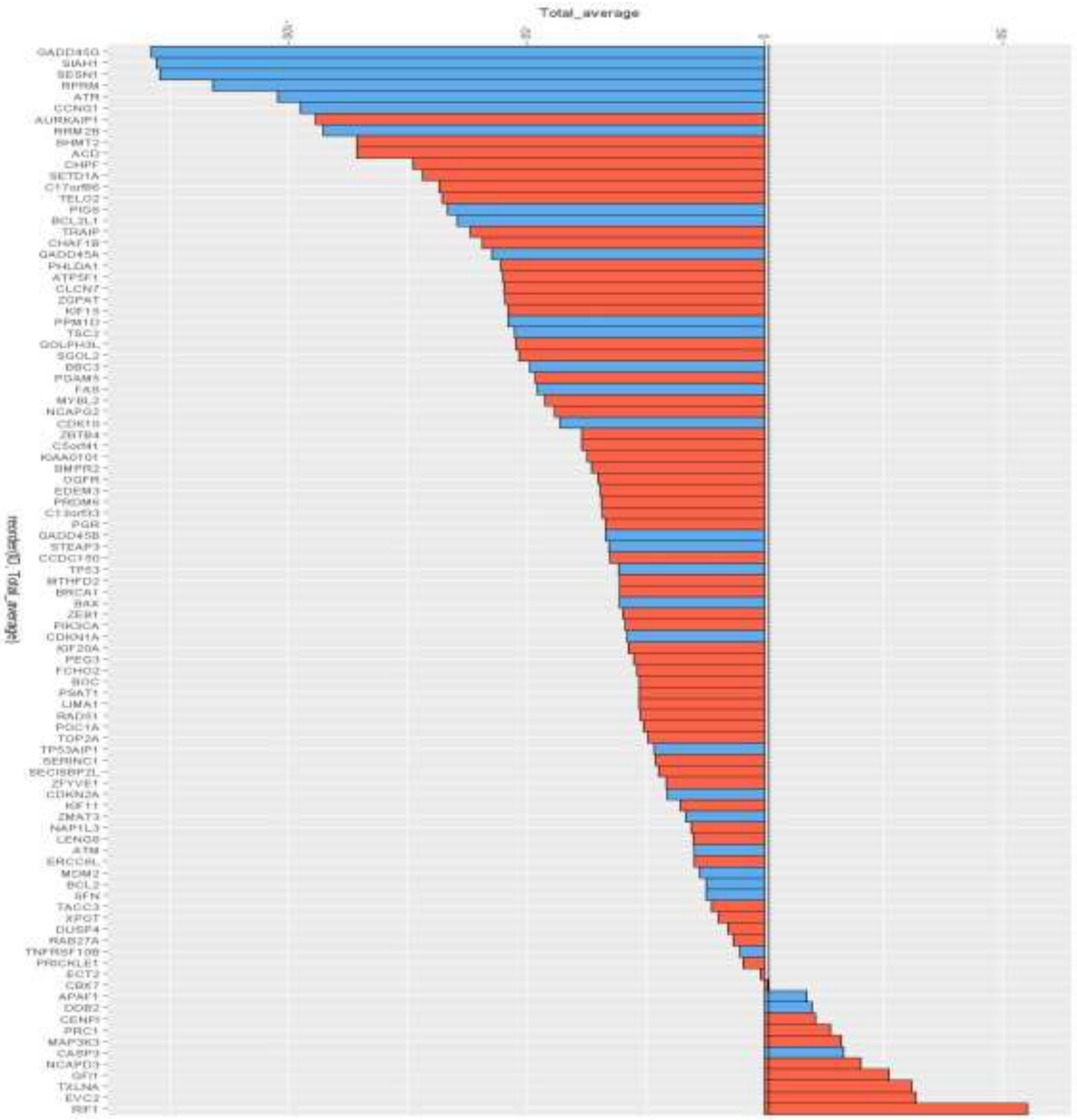


Figure 4.5: Differential drivers (as targets) associated with the TCGA-COADREAD-MissenseTP53

Sorted based on the total average of interactions. Blue colour indicates known TP53 pathway members, and red colour indicates novel targets.

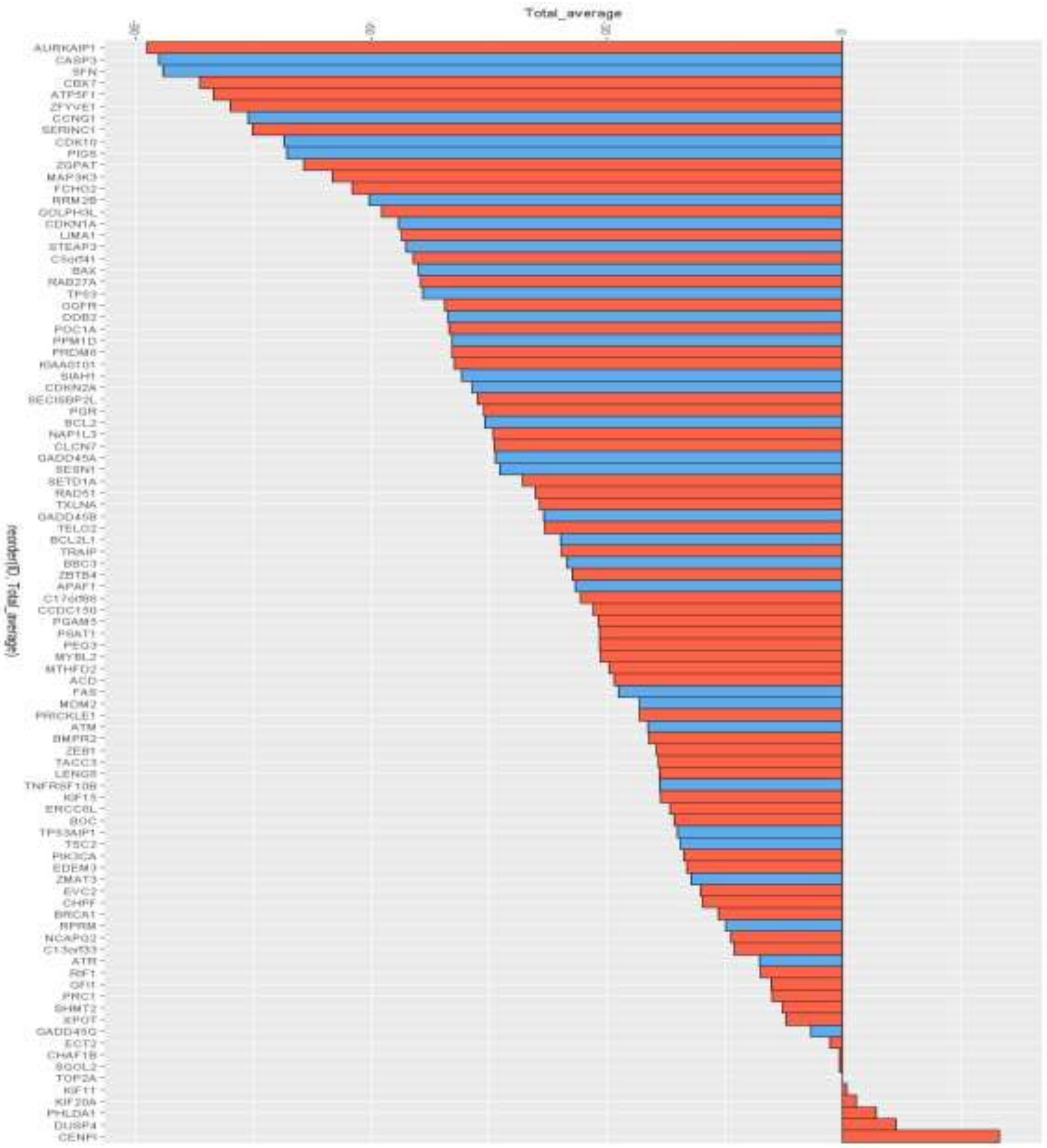


Figure 4.6: Differential drivers (as sources) associated with the TCGA-COADREADMissenseTP53

Sorted based on the average of interactions. Blue colour indicates known TP53 pathway members, and red colour indicates novel sources.

Further analysis was conducted using the MetaCore pathway analysis tool. The distinctive predictors for the TCGACOADREAD-MissenseTP53 cohort were submitted to a web-based platform (https://portal.genego.com/) to uncover the significant interactions; the results were then compared to the ANN driver analysis results. Although this tool uses a different statistical method based on the FDR-adjusted p-value, there was notable consistency with the ANN-driver analysis results. Specifically, four genes emerged as top differential source drivers also appeared as significant network objects (DUSP4, PHLDA1, KIF20A, and TOP2A) (refer to figure 4.6). Dual-specificity phosphatase 4 (DUSP4) has been reported to be involved in proliferation, downregulation of DUSP4 suppress the proliferation of cancer cell line (Ratsada et al., 2020). A recent publication indicates Pleckstrin homeolike domin, family A, member 1 (PHLDA1) as a TP53 target which contribute to cell apoptosis mediated by the TP53 gene (Song et al., 2023). Topoisomerase II α (TOP2A) has been identified to facilitate the development of high-grade serous ovarian cancer (Gao et al., 2020).

Additionally, a similar number of top differential target drivers (GIFI1, CASP3, MAP3K3, and DDB2) (refer to figure 4.5) were also observed as corresponding significant objects in MetaCore interaction results.

This convergence underscores the robustness of the findings across different analytical methodologies. The results are presented in Table 4.2.

Table 4.2: MetaCore Interactome results for the most significant interaction that match ANN driver analysis results

Input IDs	Network object name	Input IDs for corresponding object name	Corresponding network object name	Input IDs	Network object name	Input IDs for corresponding object name	Corresponding network object name
DUSP4	Dual Specificity Phosphatase 4	MDM2	MDM2	BAX	Bax	GFI1	Growth Factor Independent 1 Transcriptional Repressor
DUSP4	Dual Specificity Phosphatase 4	SETD1A	SET1A	MDM2	MDM2	CASP3	Caspase-3
DUSP4	Dual Specificity Phosphatase 4	TP53	p53	RAD51	Rad51	CASP3	Caspase-3
PHLDA1	Pleckstrin Homology Like Domain Family A Member 1	BAX	Bax	APAF1	Apaf-1	CASP3	Caspase-3
PHLDA1	Pleckstrin Homology Like Domain Family A Member 1	SETD1A	SET1A	ATM	ATM	CASP3	Caspase-3
PHLDA1	Pleckstrin Homology Like Domain Family A Member 1	TP53	p53	BCL2	Bcl-2	CASP3	Caspase-3
PHLDA1	Pleckstrin Homology Like Domain Family A Member 1	ZEB1	TCF8	BRCA1	Brca1	CASP3	Caspase-3
KIF20A	Kinesin Family Member 20A	MYBL2	b-Myb	FAS	FASN	MAP3K3	Mitogen-Activated Protein Kinase Kinase Kinase 3
TOP2A	DNA Topoisomerase II Alpha	ATM	ATM	BRCA1	Brca1	MAP3K3	Mitogen-Activated Protein Kinase Kinase Kinase 3

Table 4.2: MetaCore Interactome results for the most significant interaction that match ANN driver analysis results

Input IDs	Network object name	Input IDs for corresponding object name	Corresponding network object name	Input IDs	Network object name	Input IDs for corresponding object name	Corresponding network object name
TOP2A	DNA Topoisomerase II Alpha	BRCA1	Brca1	RAD51	Rad51	DDB2	Damage Specific DNA Binding Protein 2
TOP2A	DNA Topoisomerase II Alpha	MYBL2	b-Myb	TP53	p53	DDB2	Damage Specific DNA Binding Protein 2
TOP2A	DNA Topoisomerase II Alpha	PGR	PR (nuclear)	BCL2	Bcl-2	DDB2	Damage Specific DNA Binding Protein 2
TOP2A	DNA Topoisomerase II Alpha	TP53	p53	FAS	FASN	DDB2	Damage Specific DNA Binding Protein 2
				GFI1	GFI-1	DDB2	Damage Specific DNA Binding Protein 2

Blue colour indicates known TP53 pathway members, and orange colour indicates input that are novel to the TP53 pathway.

4.5.2. Analysis of TCGA-PAAD Cohort

The analysis for the TCGA-PAAD project was performed using a similar method to that described above (section 4.5.1). The results revealed 159 distinctive significant predictors for the TCGA-PAAD Wild-type and 100 for the TCGA-PAAD Mutant TP53 cohorts. The results also indicated that 92 genes are significant concordant predictors between cohorts. Table 4.3 summarizes the results on the top-ranked predictors, with their frequency of appearance and the associated p-value. The top-ranked distinctive predictors were then used to build the network inference for the TCGA-PAAD-MissenseTP53 cohort, using patient samples with TP53 Missense mutations (n = 63). A matrix of ((135x(135-1))17822) was generated from this analysis; the results were filtered out by taking the average of 10 repeats, then driver analysis was performed using the method described previously. Rank order was then applied according to the sum of the average values; the genes that got the highest sum of average values appeared at the top of the list, and so on.

The results of the top-100 interactions are presented as a Cytoscape image in Figure 4.7, showing (CDK6, PPM1D, CDKN2A, and IGFBP3) as major hub nodes for the TCGA-PAADMissenseTP53 cohort. CDKN2A appears as the main subnetwork with 13 interactions with genes unrelated to the pathway and 15 interactions with known pathway members, including CCNE1 and TNFRSF10B. CDKN2A protein is highly expressed in the Missense compared to the Wild-type cohort based on the cBioPortal result.

Table 4.3: Top-ranked predictors obtained using Stepwise ANN approach. Shows top-ranked predictors with frequency of gene appearance among all TP53 pathway members for the TCGA-PAAD (Mutant- and Wild-type TP53 cohorts.

Concordant predictors for both cohorts		Distinctive predictors for the TCGA-PAAD Mutant TP53			Distinctive predictors for the TCGA-PAAD Wildtype TP53		
Gene Symbol	P-value	Gene Symbol	Frequency	P-value	Gene Symbol	Frequency	P-value
E2F1	0.00473	EBF1	8	0.02862	S100A16	12	8.69E-05
KIAA0101	0.00025	KIF18B	8	0.00033	EFNA4	9	0.00047
ORC6L	6.17E-05	OIP5	8	1.45E-06	OSBPL3	9	8.79E-05
REV3L	0.00944	SPAG5	8	0.00142	S100A11	9	1.45E-05

Table 4.3: Top-ranked predictors obtained using Stepwise ANN approach. Shows top-ranked predictors with frequency of gene appearance among all TP53 pathway members for the TCGA-PAAD (Mutant- and Wild-type TP53 cohorts.

Concordant predictors for both cohorts		Distinctive predictors for the TCGA-PAAD Mutant TP53			Distinctive predictors for the TCGA-PAAD Wildtype TP53		
Gene Symbol	P-value	Gene Symbol	Frequency	P-value	Gene Symbol	Frequency	P-value
TPX2	1.92E-05	BUB1	7	5.42E-05	TMEM92	9	0.00018
ZWINT	7.26E-05	C1orf13 5	7	1.11E-05	ALPK1	8	0.01252
CDC20	0.0005 0	CDK1	7	1.62E-05	ANXA11	8	0.00018
CDC45	0.0031	CDKN3	7	0.0032	C19orf33	8	3.51E-05
CDC6	0.0010	CENPN	7	0.0055 2	FNDC3A	8	0.00057
FAM72 B	0.0001	DTYMK	7	0.0004 1	KLF5	8	0.00359
ITGB4	0.0003	FAM72D	7	0.0001 7	PLEK2	8	4.55E-06
MYBL2	0.0001	FAM83H	7	0.0003 5	S100A6	8	0.00033
UBE2C	1.6E-05	GTSE1	7	0.0068 5	TRIM16	8	0.00413
ANLN	1.22E-0	MAD2L1	7	0.0021 0	ATP2B1	7	0.00073
ASF1B	0.0003	MCM10	7	4.33E-05	C6orf132	7	0.00070
AURKA	0.0011	MCM4	7	0.0012 2	CARD6	7	0.03451
AURKB	0.0360	PKMYT1	7	0.0005 5	CMTM7	7	0.00020
BIRC5	0.0016	POLQ	7	0.0021 3	DEPDC1B	7	0.02724
C17orf5 3	0.0008	RACGA	7	0.0018	E2F8	7	0.00232

Table 4.3: Top-ranked predictors obtained using Stepwise ANN approach. Shows top-ranked predictors with frequency of gene appearance among all TP53 pathway members for the TCGA-PAAD (Mutant- and Wild-type TP53 cohorts.

Concordant predictors for both cohorts		Distinctive predictors for the TCGA-PAAD Mutant TP53			Distinctive predictors for the TCGA-PAAD Wildtype TP53		
Gene Symbol	P-value	Gene Symbol	Frequency	P-value	Gene Symbol	Frequency	P-value
		P1		9			
DTL	*0.0094*	*SMAD5*	*7*	*0.00118*	*ECT2*	*7*	*0.00031*
EPR1	*0.0011*	*SOCS2*	*7*	*8.27E-06*	*ERBB2*	*7*	*8.99E-05*
EXO1	*0.0044*	*TACC3*	*7*	*0.01188*	*FHL2*	*7*	*0.00031*
HJURP	*2.25E-0*	*TNRC6C*	*7*	*1.75E-05*	*FRRS1*	*7*	*0.00314*
KIF15	*0.0002*	*UBE2T*	*7*	*0.00324*	*GBP2*	*7*	*0.0015*

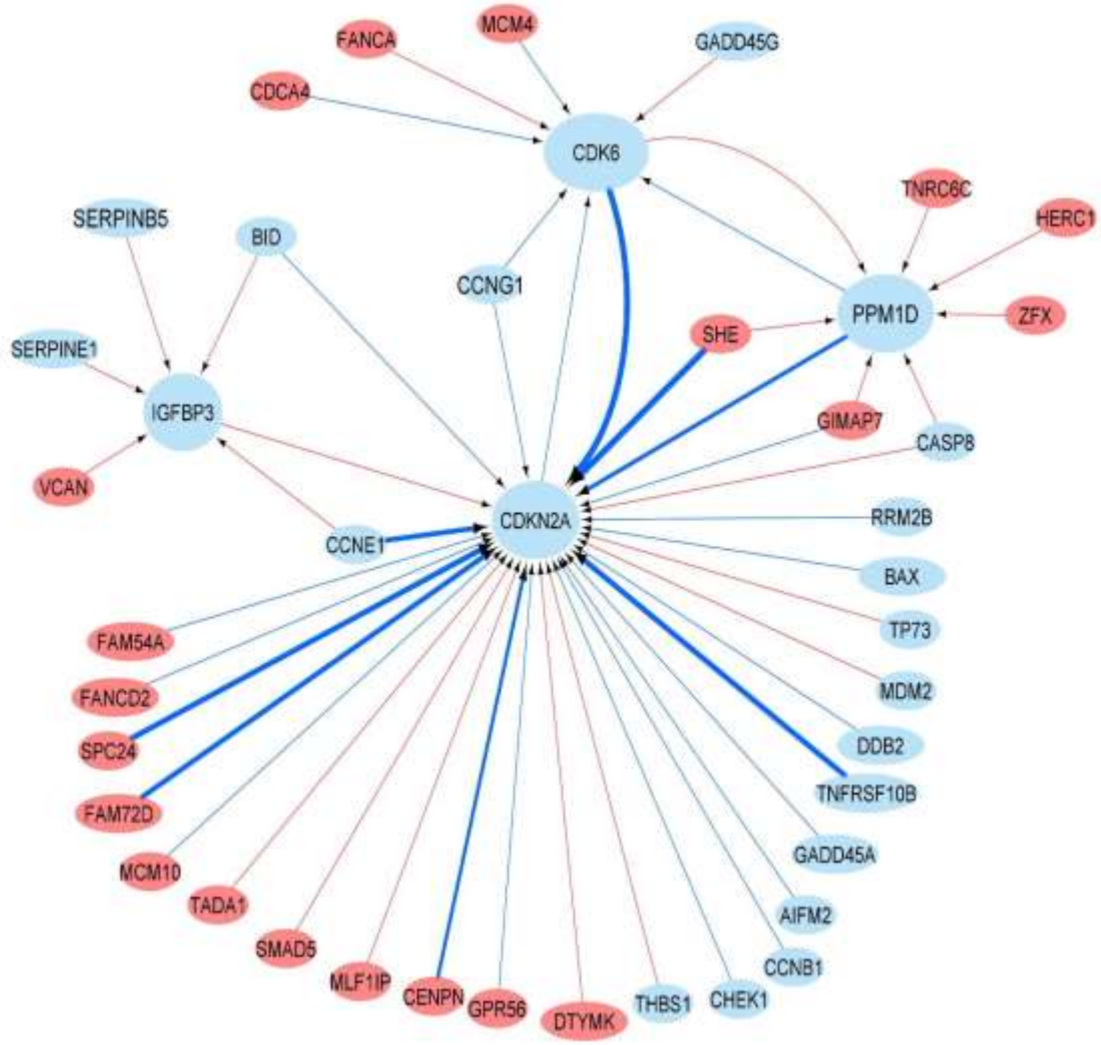


Figure 4.7: Cytoscape image for the TCGA-PAAD-MissenseTP53 cohort

The blue nodes indicate known TP53 targets, and the red nodes indicate new candidates.

The blue lines indicate positive interactions, and the red lines indicate negative interactions. Line thickness indicates interaction strength. Hub nodes are those with more than 5 interactions.

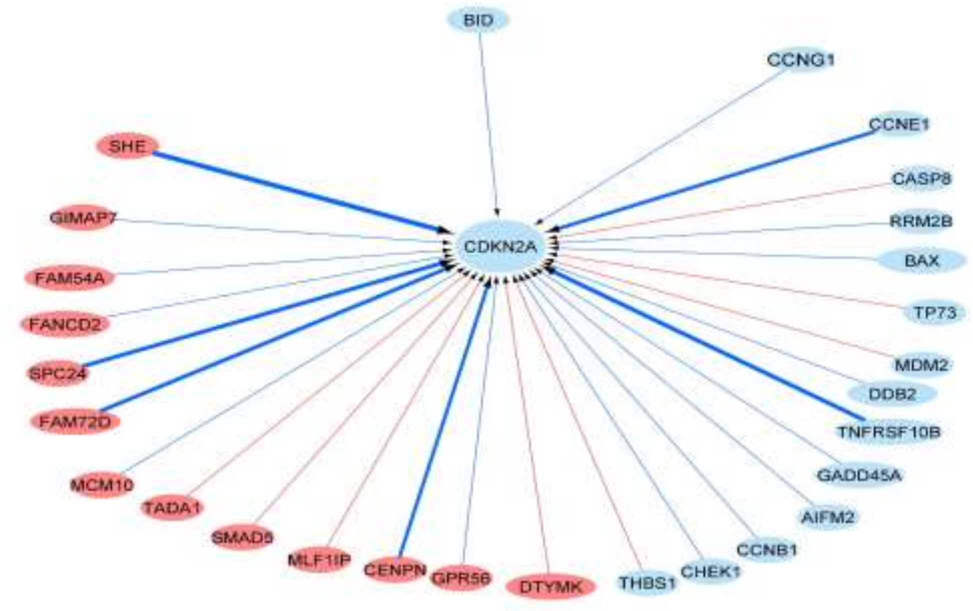


Figure 4.8: TCGA-PAAD-MissenseTP53-CDKN2A subnetwork

Contains 15 interactions with known pathway members and 13 interactions with genes not belonging to the pathway.

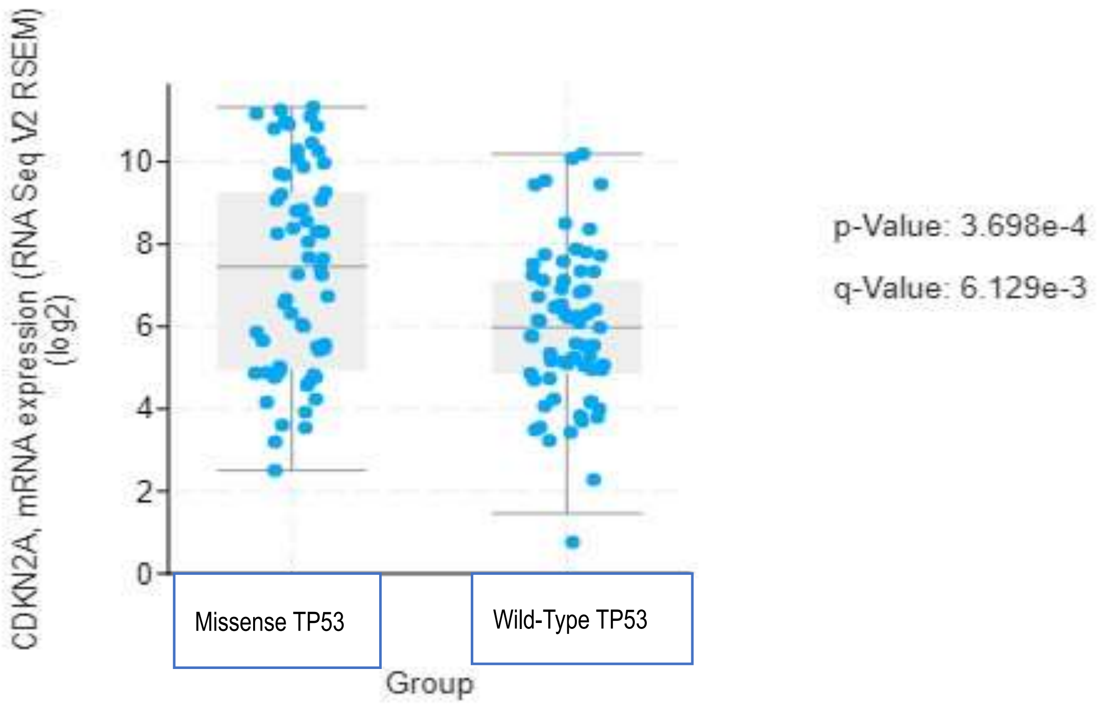


Figure 4.9: Expression of CDKN2A gene in cBioPortal

The gene is more highly expressed in the TCGA-PAAD-MissenseTP53 cohort compared to the TCGA-PAAD Wild-type TP53 cohort

Source: cBioPortal

Graphs representing the differential drivers (as targets and sources) for the TCGA-PAADMissenseTP53 cohort were made for each group based on the method described previously. Figure 4.10 shows differential drivers (as targets); 48 out of 90 got total positive interactions, including POLE2, UBE2T, FANCA, POLQ, and PLK4, which appeared as the top targets. The remaining 42 have a general negative influence. Figure 4.11 indicates differential drivers (as sources), showing that 43 out of 90 had positive interactions, 10 of which are known pathway members. The remaining 47 have a general negative influence; 23 of them are known pathway members.

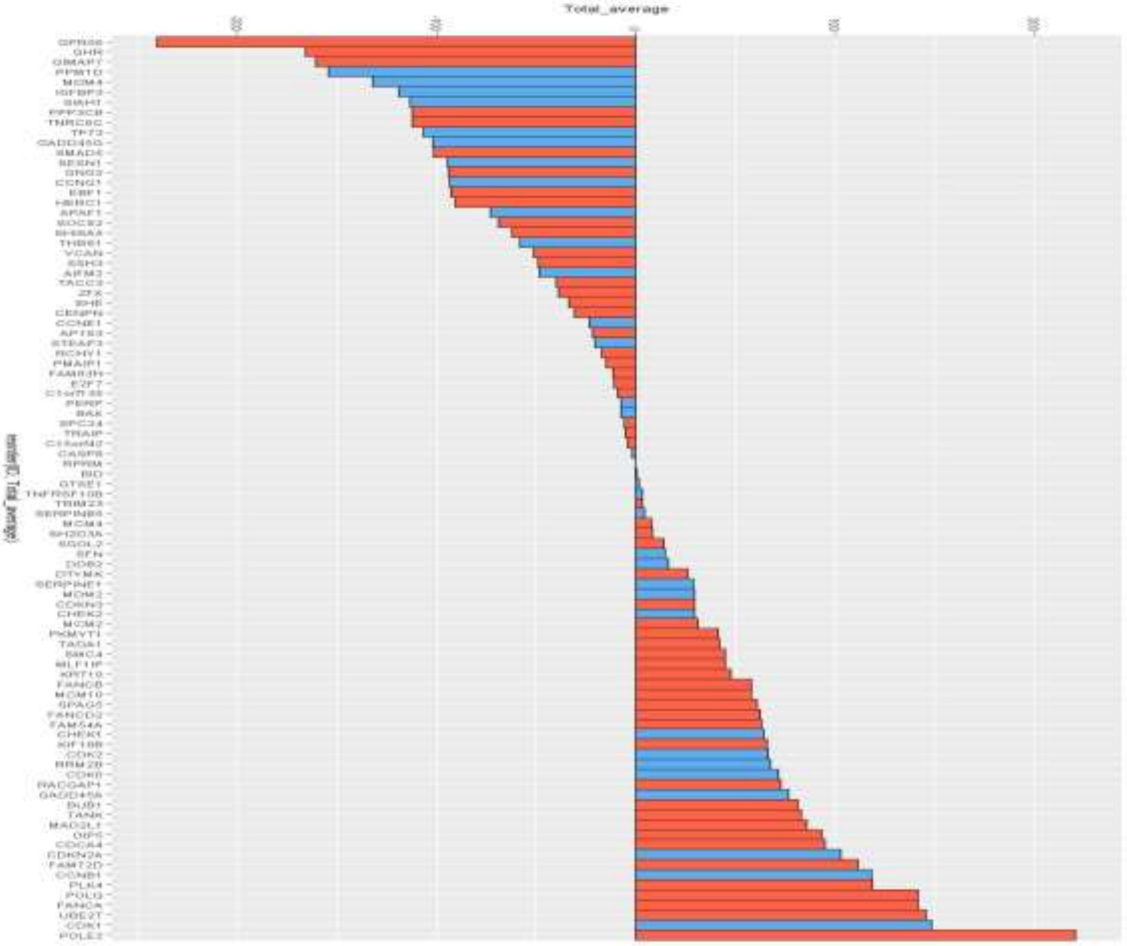


Figure 4.10: Differential drivers (as targets) associated with the TCGA-PAAD-MissenseTP53

Sorted based on the total average of interactions. Blue colour indicates known TP53 pathway members, and red colour indicates novel targets

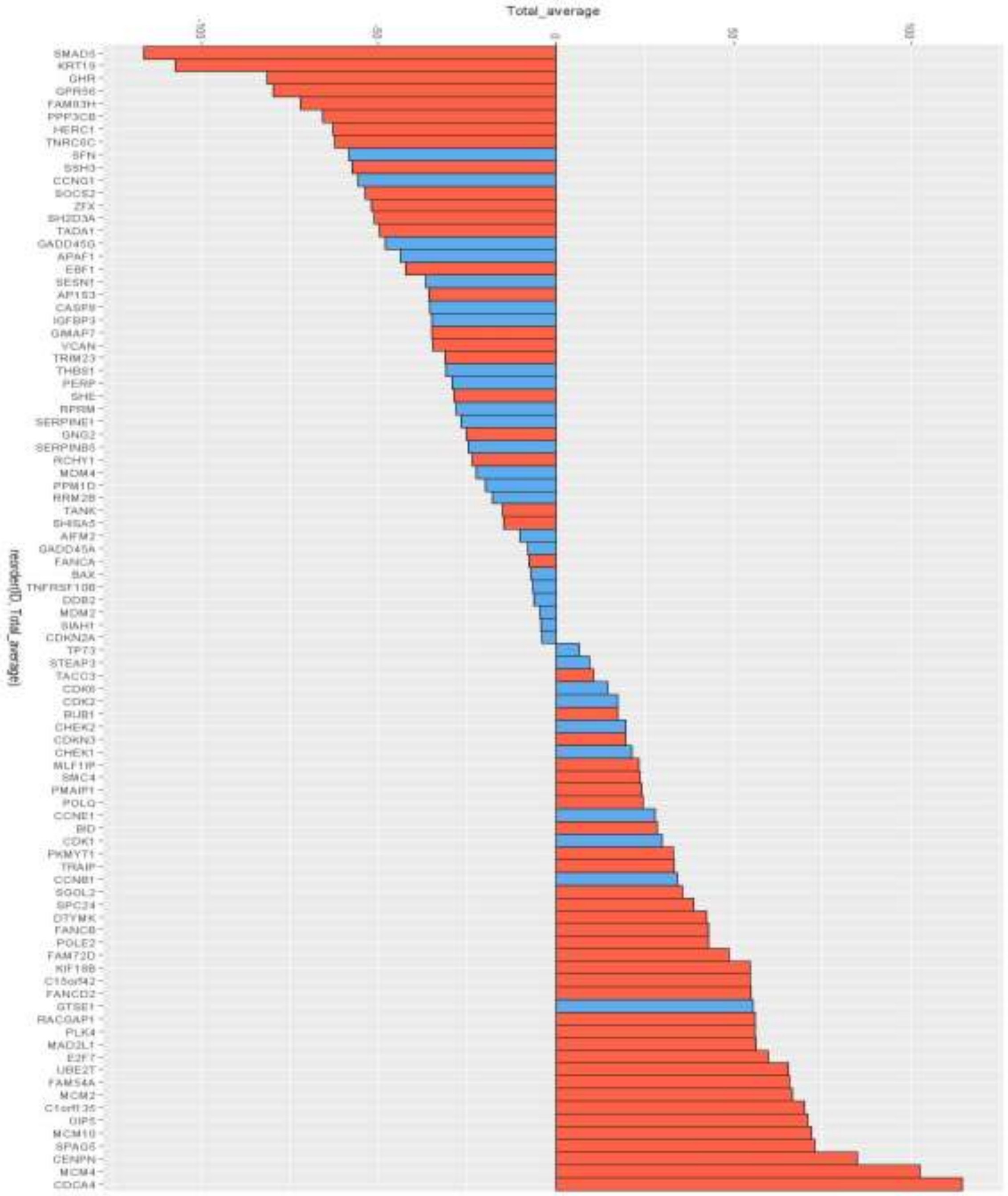


Figure 4.11: Differential drivers (as sources) associated with the TCGA-PAADMissenseTP53 Sorted based on the average of interactions. Blue colour indicates known TP53 pathway members, and red colour indicates novel sources.

MetaCore Interactome analysis was also done for this set, and the results indicate nine significant network objects (C15orf42, CENPN, E2F7, FANCD2, GTSE1, MCM10, MCM2, MCM4 and SPAG5) which appeared among the top differential sources in ANN driver analysis results. Microchromosome maintenance (MCM) proteins, a group of nuclear proteins that play important roles in cancer development by impacting cellular DNA replication. MCM10 is essential for preserving and elongating DNA replication and it is notably overproduced in various cancer tissues, thereby regulating the biological behaviour of cancer cells. MCM10 has been proposed as a predictive and diagnostic biomarker for immunomodulation as well as a possible target for tumour therapy (Chen et al., 2023). Moreover, eight corresponding significant objects (CDK1, CDK2, CDK6, Cycline B, FANCA, GADD45A, RRM2B, and TANK) appeared among the top target drivers in the ANN driver analysis outcomes. Table 4.4 highlighting the relevance and interconnectedness of these key molecular players in the context of this cohort.

Table 4.4: MetaCore Interactome results for the most significant interaction for the TCGA-PAAD-MissenseTP53 cohort matching ANN driver analysis results.

Input IDs	Network object name	Input IDs for corresponding object name	Corresponding network object name	Input IDs	Network object name	Corresponding network object name	Input IDs for corresponding object name
Orange colour indicates novel inputs and blue colour indicates known TP53 pathway members. Orange colour indicates novel inputs and blue colour indicates known TP53 pathway members							
C15orf42	TOPBP1 interacting checkpoint and replication regulator	CDK1	CDK1 (p34)	BAX	Bax	CDK1 (p34)	Cyclin dependent kinase 1
C15orf42	TOPBP1 interacting checkpoint and replication regulator	CDK2	CDK2	BCL2	Bcl-2	CDK1 (p34)	Cyclin dependent kinase 1
C15orf42	TOPBP1 interacting checkpoint and replication regulator	CHEK1	Chk1	BCL2L1	Bcl-XL	CDK1 (p34)	Cyclin dependent kinase 1
C15orf42	TOPBP1 interacting checkpoint and replication regulator	ZFX	ZFX	BRCA1	Brca1	CDK1 (p34)	Cyclin dependent kinase 1
CENPN	centromere protein N	CENPI	CENP-I (FSHPRH1)	BUB1	BUB1	CDK1 (p34)	Cyclin dependent kinase 1
E2F7	E2F transcription factor 7	CCNB1	Cyclin B1	C15orf42	Treslin	CDK1 (p34)	Cyclin dependent kinase 1

Table 4.4: MetaCore Interactome results for the most significant interaction for the TCGA-PAAD-MissenseTP53 cohort matching ANN driver analysis results.

Input IDs	Network object name	Input IDs for corresponding object name	Corresponding network object name	Input IDs	Network object name	Corresponding network object name	Input IDs for corresponding object name
E2F7	E2F transcription factor 7	CCNE1	Cyclin E	CCNB1	Cyclin B1	CDK1 (p34)	Cyclin dependent kinase 1
E2F7	E2F transcription factor 7	CDK1	CDK1 (p34)	CHAF1B	ChAF1 subunit B	CDK1 (p34)	Cyclin dependent kinase 1
E2F7	E2F transcription factor 7	CDK2	CDK2	CHEK2	Chk2	CDK1 (p34)	Cyclin dependent kinase 1
E2F7	E2F transcription factor 7	CDKN1A	p21	E2F7	E2F7	CDK1 (p34)	Cyclin dependent kinase 1
E2F7	E2F transcription factor 7	CHAF1B	ChAF1 subunit B	FANCA	FANCA	CDK1 (p34)	Cyclin dependent kinase 1
E2F7	E2F transcription factor 7	CHEK1	Chk1	MCM2	MCM7	CDK1 (p34)	Cyclin dependent kinase 1
E2F7	E2F transcription factor 7	MAD2L1	MAD2a	MCM2	MCM2	CDK1 (p34)	Cyclin dependent kinase 1
E2F7	E2F transcription factor 7	MCM2	MCM7	MCM4	MCM4	CDK1 (p34)	Cyclin dependent kinase 1
E2F7	E2F transcription factor 7	MCM2	MCM2	NCAPD3	NCAPD3	CDK1 (p34)	Cyclin dependent kinase 1
E2F7	E2F transcription factor 7	MCM4	MCM4	OGFR	OGFR	CDK1 (p34)	Cyclin dependent kinase 1

Table 4.4: MetaCore Interactome results for the most significant interaction for the TCGA-PAAD-MissenseTP53 cohort matching ANN driver analysis results.

Input IDs	Network object name	Input IDs for corresponding object name	Corresponding network object name	Input IDs	Network object name	Corresponding network object name	Input IDs for corresponding object name
E2F7	E2F transcription factor 7	MLF1IP	CENP-50	PRC1	PRC1	CDK1 (p34)	Cyclin dependent kinase 1
E2F7	E2F transcription factor 7	PRC1	PRC1	SERPINB5	Maspin	CDK1 (p34)	Cyclin dependent kinase 1
E2F7	E2F transcription factor 7	RAD51	Rad51	SFN	14-3-3 sigma	CDK1 (p34)	Cyclin dependent kinase 1
E2F7	E2F transcription factor 7	SERPINE1	PAI1	SPAG5	DEEPEST	CDK1 (p34)	Cyclin dependent kinase 1
E2F7	E2F transcription factor 7	TOP2A	TOP2 alpha	TOP2A	TOP2 alpha	CDK1 (p34)	Cyclin dependent kinase 1
E2F7	E2F transcription factor 7	TP53	p53	ATP5F1	ATP5F1	CDK2	Cyclin dependent kinase 2
E2F7	E2F transcription factor 7	UBE2T	UBE2T	BCL2	Bcl-2	CDK2	Cyclin dependent kinase 2
FANCD2	FA complementation group D2	CHEK1	Chk1	BRCA1	Brca1	CDK2	Cyclin dependent kinase 2
GTSE1	G2 and S-phase expressed 1	CDKN1A	p21	CBX7	CBX7	CDK2	Cyclin dependent kinase 2
GTSE1	G2 and S-phase expressed 1	DDB2	DDB2	CCNB1	Cyclin B1	CDK2	Cyclin dependent kinase 2

Table 4.4: MetaCore Interactome results for the most significant interaction for the TCGA-PAAD-MissenseTP53 cohort matching ANN driver analysis results.

Input IDs	Network object name	Input IDs for corresponding object name	Corresponding network object name	Input IDs	Network object name	Corresponding network object name	Input IDs for corresponding object name
GTSE1	G2 and S-phase expressed 1	TACC3	TACC3	CCNE1	LMW-CCNE1	CDK2	Cyclin dependent kinase 2
GTSE1	G2 and S-phase expressed 1	TP53	p53	CDK1	CDK1 (p34)	CDK2	Cyclin dependent kinase 2
MCM10	minichromosome maintenance 10 replication initiation factor	CDK6	CDK6	CHAF1B	ChAF1 subunit B	CDK2	Cyclin dependent kinase 2
MCM10	minichromosome maintenance 10 replication initiation factor	CDKN1A	p21	E2F7	E2F7	CDK2	Cyclin dependent kinase 2
MCM10	minichromosome maintenance 10 replication initiation factor	MCM2	MCM7	KIAA0101	p15(PAF)	CDK2	Cyclin dependent kinase 2
MCM10	minichromosome maintenance 10 replication initiation factor	MCM2	MCM2	MCM2	MCM7	CDK2	Cyclin dependent kinase 2
MCM10	minichromosome maintenance 10 replication initiation factor	MCM4	MCM4	MCM2	MCM2	CDK2	Cyclin dependent kinase 2
MCM2	minichromosome maintenance complex component 2	ATM	ATM	MCM4	MCM4	CDK2	Cyclin dependent kinase 2
MCM2	minichromosome maintenance complex component 2	ATR	ATR	PRC1	PRC1	CDK2	Cyclin dependent kinase 2
MCM2	minichromosome maintenance complex component 2	ATR	ATR	PSAT1	PSAT	CDK2	Cyclin dependent kinase 2

Table 4.4: MetaCore Interactome results for the most significant interaction for the TCGA-PAAD-MissenseTP53 cohort matching ANN driver analysis results.

Input IDs	Network object name	Input IDs for corresponding object name	Corresponding network object name	Input IDs	Network object name	Corresponding network object name	Input IDs for corresponding object name
MCM2	minichromosome maintenance complex component 2	CCNB1	Cyclin B1	SFN	14-3-3 sigma	CDK2	Cyclin dependent kinase 2
MCM2	minichromosome maintenance complex component 7	CCNE1	Cyclin E	SMC4	CAP-C	CDK2	Cyclin dependent kinase 2
MCM2	minichromosome maintenance complex component 2	CCNE1	Cyclin E	BCL2	Bcl-2	CDK6	Cyclin dependent kinase 6
MCM2	minichromosome maintenance complex component 2	CDK1	CDK1 (p34)	CDKN2A	p14ARF	CDK6	Cyclin dependent kinase 6
MCM2	minichromosome maintenance complex component 2	CDK1	CDK1 (p34)	MCM10	MCM10	CDK6	Cyclin dependent kinase 6
MCM2	minichromosome maintenance complex component 2	CDK2	CDK2	MCM2	MCM2	CDK6	Cyclin dependent kinase 6
MCM2	minichromosome maintenance complex component 2	CDK2	CDK2	RCHY1	PIRH2	CDK6	Cyclin dependent kinase 6
MCM2	minichromosome maintenance complex component 2	CDK6	CDK6	SETD1A	SET1A	CDK6	Cyclin dependent kinase 6
MCM2	minichromosome maintenance complex component 2	CDKN1A	p21	E2F7	E2F7	Cyclin B1	Cyclin B1
MCM2	minichromosome maintenance complex component 2	CHEK1	Chk1	KIAA0101	p15(PAF)	Cyclin B1	Cyclin B1

Table 4.4: MetaCore Interactome results for the most significant interaction for the TCGA-PAAD-MissenseTP53 cohort matching ANN driver analysis results.

Input IDs	Network object name	Input IDs for corresponding object name	Corresponding network object name	Input IDs	Network object name	Corresponding network object name	Input IDs for corresponding object name
MCM2	minichromosome maintenance complex component 2	CHEK1	Chk1	MCM2	MCM7	Cyclin B1	Cyclin B1
MCM2	minichromosome maintenance complex component 2	DDB2	DDB2	FANCB	FANCB	FANCA	FA complementation group A
MCM2	minichromosome maintenance complex component 2	FANCD2	FANCD2	BRCA1	Brca1	GADD45 alpha	Growth arrest and DNA damage inducible alpha
MCM2	minichromosome maintenance complex component 2	FANCD2	FANCD2	CCNB1	Cyclin B1	GADD45 alpha	Growth arrest and DNA damage inducible alpha
MCM2	minichromosome maintenance complex component 2	MCM2	MCM2	CDK1	CDK1 (p34)	GADD45 alpha	Growth arrest and DNA damage inducible alpha
MCM2	minichromosome maintenance complex component 2	MCM4	MCM4	FAS	FasR(CD95)	GADD45 alpha	Growth arrest and DNA damage inducible alpha
MCM2	minichromosome maintenance complex component 2	PGR	PR (nuclear)	GADD45B	GADD45 beta	GADD45 alpha	Growth arrest and DNA damage inducible alpha
MCM2	minichromosome maintenance complex component 2	RAD51	Rad51	MDM2	MDM2	GADD45 alpha	Growth arrest and DNA damage

Table 4.4: MetaCore Interactome results for the most significant interaction for the TCGA-PAAD-MissenseTP53 cohort matching ANN driver analysis results.

Input IDs	Network object name	Input IDs for corresponding object name	Corresponding network object name	Input IDs	Network object name	Corresponding network object name	Input IDs for corresponding object name
							inducible alpha
MCM2	minichromosome maintenance complex component 7	TP53	p53	SETD1A	SET1A	GADD45 alpha	Growth arrest and DNA damage inducible alpha
MCM4	minichromosome maintenance complex component 4	CASP8	Caspase-8	TP53	p53	GADD45 alpha	Growth arrest and DNA damage inducible alpha
MCM4	minichromosome maintenance complex component 4	CDK1	CDK1 (p34)	TP73	p73	GADD45 alpha	Growth arrest and DNA damage inducible alpha
MCM4	minichromosome maintenance complex component 4	CDK2	CDK2	ZFX	ZFX	GADD45 alpha	Growth arrest and DNA damage inducible alpha
MCM4	minichromosome maintenance complex component 4	FANCD2	FANCD2	ATM	ATM	RRM2B	Ribonucleotide reductase regulatory TP53 inducible subunit M2B
MCM4	minichromosome maintenance complex component 4	MCM2	MCM2	BCL2	Bcl-2	RRM2B	Ribonucleotide reductase regulatory TP53 inducible subunit M2B

Table 4.4: MetaCore Interactome results for the most significant interaction for the TCGA-PAAD-MissenseTP53 cohort matching ANN driver analysis results.

Input IDs	Network object name	Input IDs for corresponding object name	Corresponding network object name	Input IDs	Network object name	Corresponding network object name	Input IDs for corresponding object name
MCM4	minichromosome maintenance complex component 4	TP53	p53	MDM2	MDM2	RRM2B	Ribonucleotide reductase regulatory TP53 inducible subunit M2B
MCM4	minichromosome maintenance complex component 4	ZFX	ZFX	TP53	p53	RRM2B	Ribonucleotide reductase regulatory TP53 inducible subunit M2B
SPAG5	sperm associated antigen 5	ATM	ATM	ZFX	ZFX	RRM2B	Ribonucleotide reductase regulatory TP53 inducible subunit M2B
SPAG5	sperm associated antigen 5	CDK1	CDK1 (p34)	CASP3	Caspase-3	TANK	TRAF family member associated NFKB activator
SPAG5	sperm associated antigen 5	PLK4	PLK4 (STK18)	CASP8	Caspase-8	TANK	TRAF family member associated NFKB activator
SPAG5	sperm associated antigen 5	ZFX	ZFX	SETD1A	SET1A	TANK	TRAF family member associated NFKB activator

4.5.3. Analysis of TCGA-STAD Cohort

The methods described previously (in sections 4.5.1 and 4.5.2) were used for the analysis of this cohort. The results showed a total of 377 distinctive significant predictors for the TCGASTAD Wild-type TP53 cohort; 241 distinctive significant predictors for the TCGA-STAD Mutant TP53 cohort; and 47 significant concordant predictors between the two cohorts. Table 4.5 summarizes the results of the top-ranked predictors, with frequency of occurrence and associated P-values for both cohorts. The top-ranked predictors were used to build the interaction network for the TCGA-STAD- MissenseTP53 cohort, which contains 123 samples with TP53 Missense mutations. This analysis produces a matrix of ((123x(123-1))111222). The average of interaction and driver analysis was performed using the method described in Chapter 3. Genes were then ordered based on the highest sum of the average values.

The results of the top-100 interactions are presented as a Cytoscape image in Figure 4.12, showing CDKN2A, MDM2, PMAIP1, BCL2L1, ZMAT5, SF3B1, and CDKN1A as major hub nodes for the TCGA-STAD-MissenseTP53 cohort. CDKN2A appeared as the main subnetwork, with nine interactions with genes not related to the pathway, and five interactions with known pathway members, including CCNE1, TP73, PMAIP1, CASP9, and GORAB. cBioPortal results also indicated higher CDKN2A protein expression in the Missense compared to the Wild-type cohort.

Table 4.5: Top-ranked predictors with frequency of gene appearance among all TP53 pathway members for the TCGA-STAD (Mutant- and Wild-type TP53 cohorts)

Concordant predictors for Both cohorts		Distinctive predictors for TCGA-STAD Mutant TP53			Distinctive predictors for the TCGA-STAD Wild TP53		
Gene Symbol	P-value	Gene Symbol	Frequency	P-value	Gene Symbol	Frequency	P-value
BUB1	1.3425E-51	EXO1	9	3.5567E-06	SNORD115-17	19	4.6176 E-39
CDCA8	7.7448E-50	CCNA2	8	3.0602E-12	BARX1	17	5.4463 E-16
CDCA3	4.1505E-41	CENPA	8	3.9839E-09	MUC13	16	2.7321 E-47
POLE2	1.0669E-49	DSCC1	8	3.6864E-33	TUBG2	16	9.6025 E-43

Table 4.5: Top-ranked predictors with frequency of gene appearance among all TP53 pathway members for the TCGA-STAD (Mutant- and Wild-type TP53 cohorts)

Concordant predictors for Both cohorts		Distinctive predictors for TCGA-STAD Mutant TP53			Distinctive predictors for the TCGA-STAD Wild TP53		
Gene Symbol	P-value	Gene Symbol	Frequency	P-value	Gene Symbol	Frequency	P-value
CENPF	1.818E-58	DTL	8	1.2968E-17	ADCY8	15	6.0706E-41
FOXM1	5.1384E-43	ERCC6L	8	1.074E-43	FERMT1	14	3.7921E-38
FAM54A	1.5191E-49	KIF18B	8	5.2841E-06	SNORD115-41	13	3.1906E-46
CDCA2	6.0297E-33	MAD2L1	8	1.5134E-10	FLJ42393	13	1.3841E-45
CASC5	4.1383E-50	ORC1L	8	2.003E-36	CNPY2	13	7.1919E-42
C12orf48	6.8986E-45	PRIM1	8	4.4905E-13	GRINA	13	1.035E-40
BUB1B	1.3428E-50	RFC3	8	0.00418791	TRIM15	12	8.3554E-47
NCAPH	3.4128E-46	RNF150	8	0.01926792	TCEA2	12	4.5471E-45
RRM2	4.1269E-35	SPAG5	8	3.2241E-15	SNORD29	12	1.8176E-33
UBE2C	4.0639E-62	TRIP13	8	2.3199E-10	LPP	12	1.5986E-32
CCNF	3.7804E-23	TROAP	8	6.0609E-05	SNORA36C	12	2.1564E-13
NCAPG	4.0629E-48	AHCTF1	7	7.4581E-17	ZNF559	11	4.8751E-

Table 4.5: Top-ranked predictors with frequency of gene appearance among all TP53 pathway members for the TCGA-STAD (Mutant- and Wild-type TP53 cohorts)

Concordant predictors for Both cohorts		Distinctive predictors for TCGA-STAD Mutant TP53			Distinctive predictors for the TCGA-STAD Wild TP53		
Gene Symbol	P-value	Gene Symbol	Frequency	P-value	Gene Symbol	Frequency	P-value
							46
CDC25C	2.2008E-29	ATP8B2	7	1.3856E-15	ESRP1	11	1.0525E-45
SKA1	2.3217E-34	AURKA	7	0.00026095	FA2H	11	2.0893E-45
CDC20	1.1016E-40	C10orf72	7	0.00046173	TRIM31	11	4.3576E-43
CDCA5	2.4148E-31	C11orf82	7	1.5658E-06	VEGFB	11	1.0008E-39
FBN1	2.8004E-11	C1orf112	7	5.4377E-06	GPR35	11	1.5679E-39
GSG2	6.5295E-41	C21orf45	7	1.9437E-19	AGR2	11	1.0475E-30
NEK2	1.6142E-36	CCNB2	7	1.7886E-15	FBLL1	11	1.967E-30
PRR11	1.8375E-44	CDC25A	7	7.3221E-30	BNIPL	11	4.7807E-11
Source: CENPO	5.2289E-44	CDC45	7	2.743E-23	CLSPN	10	6.4374E-47

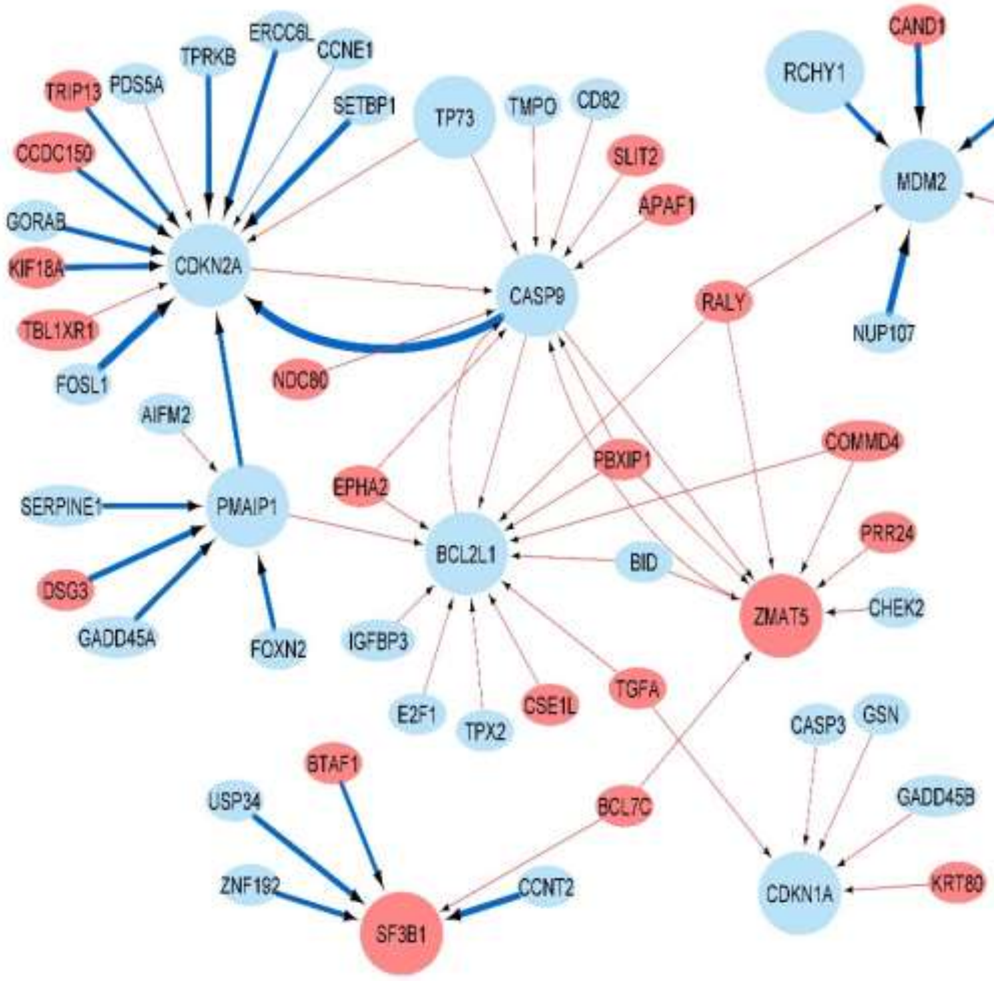


Figure 4.12: Cytoscape image for the TCGA-STAD-MissenseTP53 cohort

The blue nodes indicate known TP53 targets, and the red nodes indicate new candidates.

The blue lines indicate positive interactions, and the red lines indicate negative interactions. Line thickness indicates interaction strength. Hub nodes are those with more than 5 interactions

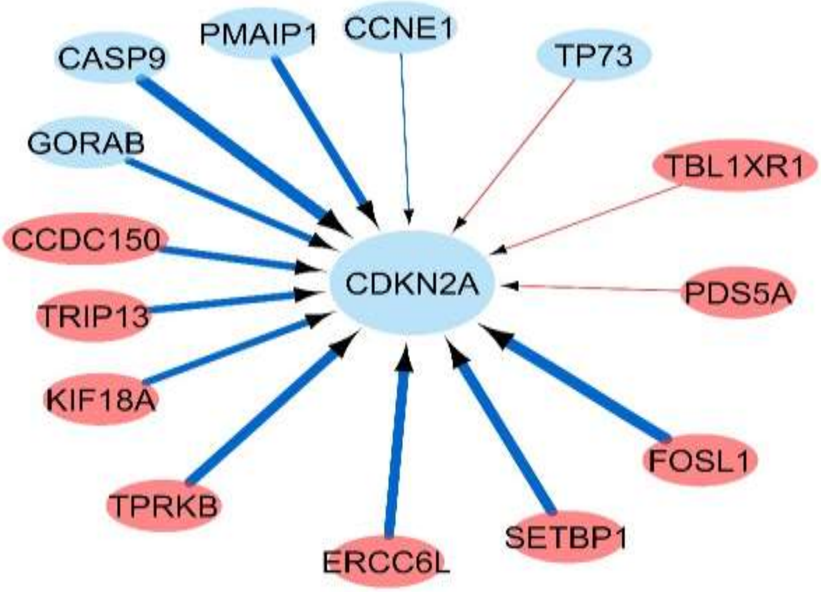


Figure 4.13: TCGA-STAD-MissenseTP53-CDKN2A subnetwork

Contains 5 interactions with known pathway members and 9 interactions with genes not belonging to the pathway

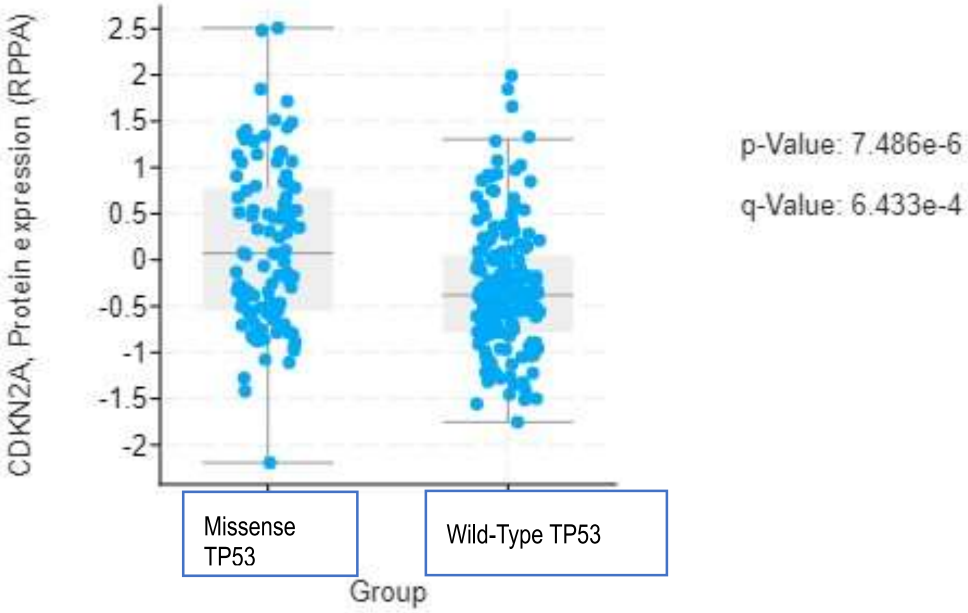


Figure 4.14: Expression of CDKN2A gene in cBioPortal

The gene is more highly expressed in the TCGA-STAD-MissenseTP53 cohort compared to the TCGA-STAD WTTP53 cohort

Source: cBioPortal

Graphs that represent the differential drivers as targets and sources were made based on the methods described previously. Figure 4.14 represents differential drivers as targets for the

TCGA-STAD-MissenseTP53 cohort. 17 out of 27 show positive interactions, among them (CDKN2A, ZNF638, PMAIP1, MDM2, and FBXN2) which appeared as top targets. The remaining 10 show negative interactions. Figure 4.15 indicates differential drivers as sources for the same cohort. 38 out of 72 with positive interaction, among them (FOXN2, USP34, NUP107, FBXO11, and CAND1), which appeared as top sources. The remaining 34 have a general negative influence.

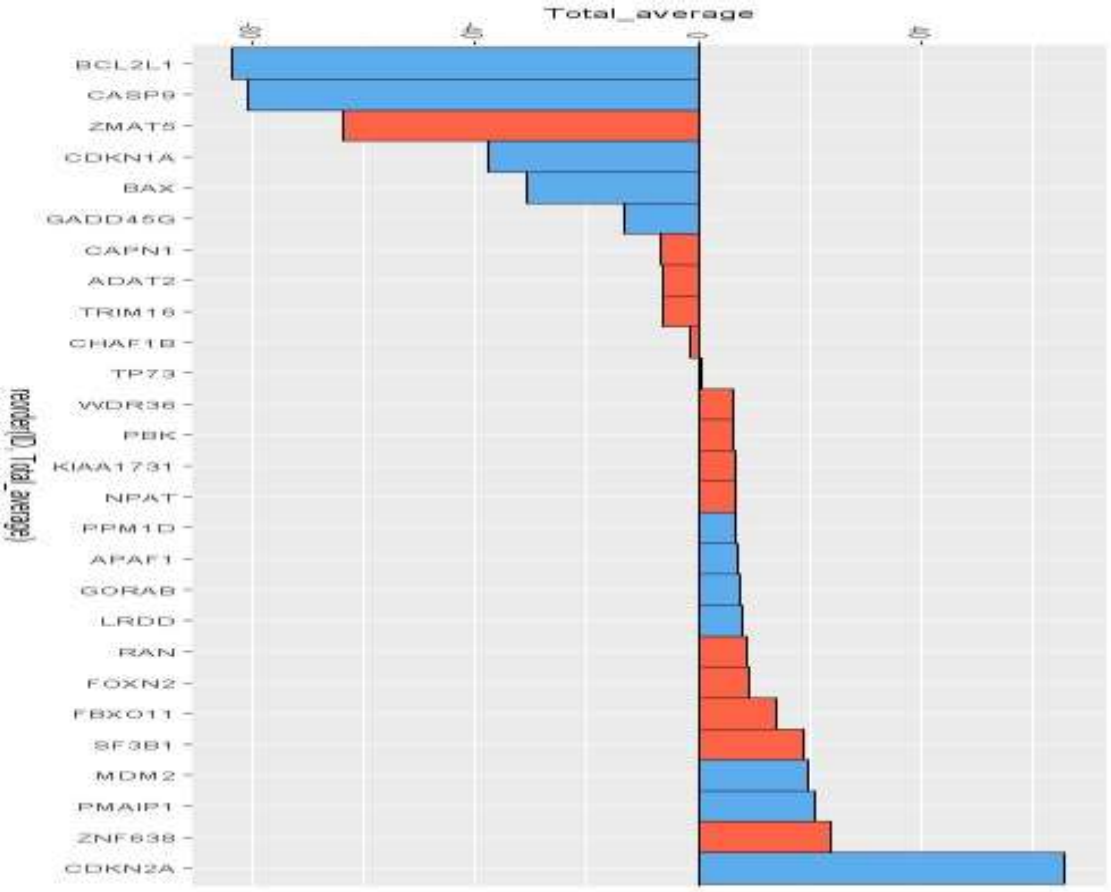


Figure 4.15: Differential drivers (as targets) associated with the TCGA-STAD-MissenseTP53 Sorted based on the total average of interactions. Blue colour indicates known TP53 pathway members, and red colour indicates novel targets

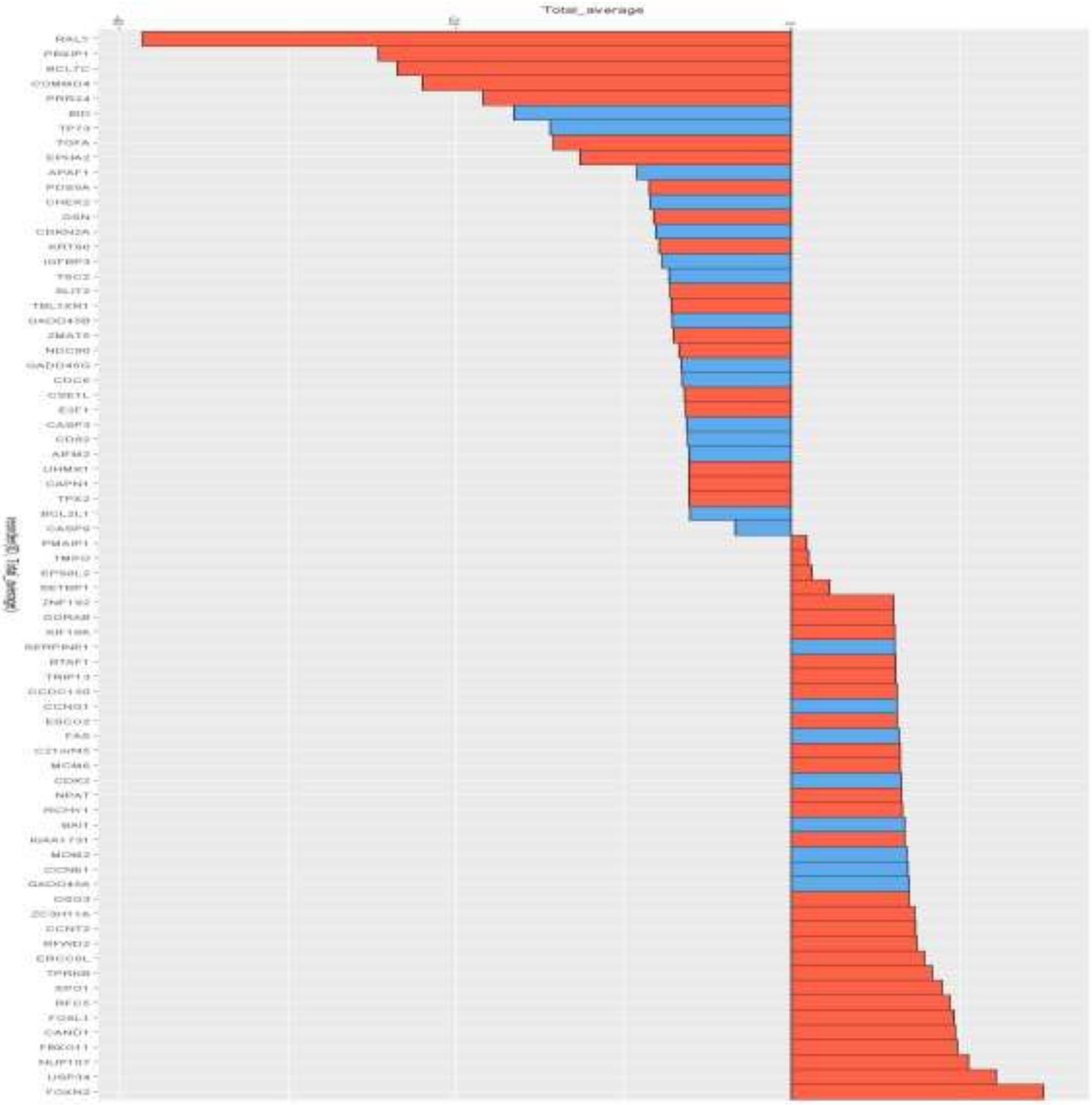


Figure 4.16: Differential drivers (As sources) associated with the TCGA-STAD-MissenseTP53 Sorted based on the average of interactions. Blue colour indicates known TP53 pathway members, and red colour indicates novel sources.

MetaCore Interactome analysis for this set indicates 13 significant network objects, which also appeared among the top differential sources in the ANN driver analysis result (CAND1, CCNE1, CCNT2, ERCC6L, FOSL1, FOXN2, GADD45A, NUP107, RFC5, RFWD2, TPRKB, XPO1, and ZC3H11A). And 12 corresponding significant objects (APAF1, CDKN2A, GORAB, LRDD, MDM2, PBK, PMAIP1, RAN, SF3B1, TP73, ZNF638) which also appeared among the top differential targets in ANN driver analysis. Table 4.6 presents the MetaCore result for this set.

Table 4.6: MetaCore Interactome results for the most significant interaction of the TCGA-STAD-MissenseTP53 cohort matching ANN driver analysis results

Input IDs	Network object name	Input IDs for correspondin g object name	Corresponding network object name	Input IDs	Network object name	Input IDs for correspondin g object name	Corresponding network object name
Orange colour indicates novel inputs and blue colour indicates known TP53 pathway members.							
CAND1	cullin associated and neddylation dissociated 1	BLM	BLM	CDC6	CDC18L (CDC6)	APAF1	apoptotic peptidase activating factor 1
CAND1	cullin associated and neddylation dissociated 1	CCNB1	Cyclin B1	CSE1L	CSE1L	APAF1	apoptotic peptidase activating factor 1
CAND1	cullin associated and neddylation dissociated 1	CDK2	CDK2	NUP107	NUP107	APAF1	apoptotic peptidase activating factor 1
CAND1	cullin associated and neddylation dissociated 1	FGFR1	FGFR1	CBX7	CBX7	CDKN2A	cyclin dependent kinase inhibitor 2A
CAND1	cullin associated and neddylation dissociated 1	MCM2	MCM2	CDC45	CDC45L	CDKN2A	cyclin dependent kinase inhibitor 2A
CAND1	cullin associated and neddylation dissociated 1	SF3B1	SF3B2	DHX9	DDX9	CDKN2A	cyclin dependent kinase inhibitor 2A

Table 4.6: MetaCore Interactome results for the most significant interaction of the TCGA-STAD-MissenseTP53 cohort matching ANN driver analysis results

Input IDs	Network object name	Input IDs for correspondin g object name	Corresponding network object name	Input IDs	Network object name	Input IDs for correspondin g object name	Corresponding network object name
CAND1	cullin associated and neddylation dissociated 1	TNRC6C	Tnrc6c	EZH2	EZH2	CDKN2A	cyclin dependent kinase inhibitor 2A
CAND1	cullin associated and neddylation dissociated 1	ZEB1	TCF8	MYSM1	MYSM1	CDKN2A	cyclin dependent kinase inhibitor 2A
CCNE1	cyclin E1	BLM	BLM	RUNX1T1	ETO	CDKN2A	cyclin dependent kinase inhibitor 2A
CCNE1	cyclin E1	CDC25A	CDC25A	SUV39H2	SUV39H2	CDKN2A	cyclin dependent kinase inhibitor 2A
CCNE1	cyclin E1	CDC25C	CDC25C	UHRF1	UHRF1	CDKN2A	cyclin dependent kinase inhibitor 2A
CCNE1	cyclin E1	CDKN1A	p21	RCHY1	PIRH2	GORAB	golgin, RAB6 interacting
CCNE1	cyclin E1	E2F1	E2F1	FANCI	FANCI (KIAA1794)	LRDD	P53-Induced Death Domain Protein 1
CCNE1	cyclin E1	FEN1	FEN1	RFC5	RFC5	LRDD	P53-Induced Death Domain Protein 1
CCNE1	cyclin E1	FOXM1	FOXM1	ATM	ATM	MDM2	MDM2 proto-oncogene
CCNE1	cyclin E1	MYBL2	b-Myb	AURKA	Aurora-A	MDM2	MDM2 proto-oncogene
CCNE1	cyclin E1	ORC1L	ORC1L	BCL2	Bcl-2	MDM2	MDM2 proto-oncogene

Table 4.6: MetaCore Interactome results for the most significant interaction of the TCGA-STAD-MissenseTP53 cohort matching ANN driver analysis results

Input IDs	Network object name	Input IDs for correspondin g object name	Corresponding network object name	Input IDs	Network object name	Input IDs for correspondin g object name	Corresponding network object name
CCNE1	cyclin E1	PTEN	PTEN	BCL2L1	Bcl-XL	MDM2	MDM2 proto-oncogene
CCNT2	cyclin T2	E2F1	E2F1	BID	Bid	MDM2	MDM2 proto-oncogene
ERCC6L	ERCC excision repair 6 like, spindle assembly checkpoint helicase	BLM	BLM	CDK2	CDK2	MDM2	MDM2 proto-oncogene
ERCC6L	ERCC excision repair 6 like, spindle assembly checkpoint helicase	CHEK1	Chk1	CDK4	CDK4	MDM2	MDM2 proto-oncogene
ERCC6L	ERCC excision repair 6 like, spindle assembly checkpoint helicase	E2F1	E2F1	CDKN2A	p14ARF	MDM2	MDM2 proto-oncogene

Table 4.6: MetaCore Interactome results for the most significant interaction of the TCGA-STAD-MissenseTP53 cohort matching ANN driver analysis results

Input IDs	Network object name	Input IDs for correspondin g object name	Corresponding network object name	Input IDs	Network object name	Input IDs for correspondin g object name	Corresponding network object name
ERCC6L	ERCC excision repair 6 like, spindle assembly checkpoint helicase	GLI3	GLI-3	CDS1; CHEK2	Chk2	MDM2	MDM2 proto-oncogene
ERCC6L	ERCC excision repair 6 like, spindle assembly checkpoint helicase	TOP2A	TOP2 alpha	DHX9	DDX9	MDM2	MDM2 proto-oncogene
FOSL1	FOS like 1, AP-1 transcription factor subunit	CCND1	Cyclin D1	DTL	DTL (hCdt2)	MDM2	MDM2 proto-oncogene
FOSL1	FOS like 1, AP-1 transcription factor subunit	FEN1	FEN1	EZH2	EZH2	MDM2	MDM2 proto-oncogene
FOSL1	FOS like 1, AP-1 transcription factor subunit	ITGA2	ITGA2	FGFR1	FGFR1	MDM2	MDM2 proto-oncogene

Table 4.6: MetaCore Interactome results for the most significant interaction of the TCGA-STAD-MissenseTP53 cohort matching ANN driver analysis results

Input IDs	Network object name	Input IDs for correspondin g object name	Corresponding network object name	Input IDs	Network object name	Input IDs for correspondin g object name	Corresponding network object name
FOSL1	FOS like 1, AP-1 transcription factor subunit	ZEB1	TCF8	MPDZ	MPDZ	MDM2	MDM2 proto-oncogene
FOXN2	forkhead box N2	E2F1	E2F1	PBXIP1	PBXIP1	MDM2	MDM2 proto-oncogene
FOXN2	forkhead box N2	FGFR1	FGFR1	PRC1	PRC1	MDM2	MDM2 proto-oncogene
FOXN2	forkhead box N2	SETD1A	SET1A	PSAT1	PSAT	MDM2	MDM2 proto-oncogene
FOXN2	forkhead box N2	ZEB1	TCF8	SETD1A	SET1A	MDM2	MDM2 proto-oncogene
GADD45 A	growth arrest and DNA damage inducible alpha	GADD45G	GADD45 gamma	SFN	14-3-3 sigma	MDM2	MDM2 proto-oncogene
NUP107	nucleoporin 107	AHCTF1	ELYS	TNFRSF10B	DR5(TNFRSF10B)	MDM2	MDM2 proto-oncogene
NUP107	nucleoporin 107	APAF1	Apaf-1	TP53	p53 (mitochondrial)	MDM2	MDM2 proto-oncogene
NUP107	nucleoporin 107	CDK2	CDK2	TP73	P73 dN-Alpha	MDM2	MDM2 proto-oncogene
NUP107	nucleoporin 107	CENPA	CENP-A	TPX2	TPX2	MDM2	MDM2 proto-oncogene

Table 4.6: MetaCore Interactome results for the most significant interaction of the TCGA-STAD-MissenseTP53 cohort matching ANN driver analysis results

Input IDs	Network object name	Input IDs for correspondin g object name	Corresponding network object name	Input IDs	Network object name	Input IDs for correspondin g object name	Corresponding network object name
NUP107	nucleoporin 107	CENPF	CENP-F	MPDZ	MPDZ	PBK	PDZ binding kinase
NUP107	nucleoporin 107	FGFR1	FGFR1	BAX	Bax	PMAIP1	phorbol-12-myristate-13-acetate-induced protein 1
NUP107	nucleoporin 107	RANBP2	RanBP2	BCL2	Bcl-2	PMAIP1	phorbol-12-myristate-13-acetate-induced protein 1
NUP107	nucleoporin 107	TNS1	CARD5	BCL2L1	Bcl-XL	PMAIP1	phorbol-12-myristate-13-acetate-induced protein 1
NUP107	nucleoporin 107	ZFX	ZFX	CSE1L	CSE1L	PMAIP1	phorbol-12-myristate-13-acetate-induced protein 1
RFC5	replication factor C subunit 5	ATM	ATM	EZH2	EZH2	PMAIP1	phorbol-12-myristate-13-acetate-induced protein 1
RFC5	replication factor C subunit 5	ATR	ATR	FGFR1	FGFR1	PMAIP1	phorbol-12-myristate-13-acetate-induced protein 1
RFC5	replication factor C subunit 5	BLM	BLM	GFI1	GFI-1	PMAIP1	phorbol-12-myristate-13-acetate-induced protein 1
RFC5	replication factor C subunit 5	CAND1	TIP120A	KPNA2	Karyopherin alpha 2	PMAIP1	phorbol-12-myristate-13-acetate-induced protein 1

Table 4.6: MetaCore Interactome results for the most significant interaction of the TCGA-STAD-MissenseTP53 cohort matching ANN driver analysis results

Input IDs	Network object name	Input IDs for correspondin g object name	Corresponding network object name	Input IDs	Network object name	Input IDs for correspondin g object name	Corresponding network object name
RFC5	replication factor C subunit 5	CCND1	Cyclin D1	MDM2	MDM2	PMAIP1	phorbol-12-myristate-13-acetate-induced protein 1
RFC5	replication factor C subunit 5	DSCC1	DCC1	SETD1A	SET1A	PMAIP1	phorbol-12-myristate-13-acetate-induced protein 1
RFC5	replication factor C subunit 5	E2F1	E2F1	TP53	p53	PMAIP1	phorbol-12-myristate-13-acetate-induced protein 1
RFC5	replication factor C subunit 5	EXOSC8	RRP43	TP73	p73	PMAIP1	phorbol-12-myristate-13-acetate-induced protein 1
RFC5	replication factor C subunit 5	FGFR1	FGFR1	CDCA8	CDCA8	PPM1D	protein phosphatase, Mg2+/Mn2+ dependent 1D
RFC5	replication factor C subunit 5	LRDD	PIDD	CDS1; CHEK2	Chk2	PPM1D	protein phosphatase, Mg2+/Mn2+ dependent 1D
RFC5	replication factor C subunit 5	RFC3	RFC3	CDCA2	CDCA2	RAN	RAN, member RAS oncogene family
RFC5	replication factor C subunit 5	RIF1	BAIP3	CDK2	CDK2	RAN	RAN, member RAS oncogene family
RFWD2	COP1 E3 Ubiquitin Ligase	MDM4	MDM4	CSE1L	CSE1L	RAN	RAN, member RAS oncogene family

Table 4.6: MetaCore Interactome results for the most significant interaction of the TCGA-STAD-MissenseTP53 cohort matching ANN driver analysis results

Input IDs	Network object name	Input IDs for correspondin g object name	Corresponding network object name	Input IDs	Network object name	Input IDs for correspondin g object name	Corresponding network object name
TPRKB	TP53RK binding protein	BIRC5	Survivin	DLGAP5	HURP	RAN	RAN, member RAS oncogene family
TPRKB	TP53RK binding protein	E2F1	E2F1	FANCA	FANCA	RAN	RAN, member RAS oncogene family
TPRKB	TP53RK binding protein	FGFR1	FGFR1	FGFR1	FGFR1	RAN	RAN, member RAS oncogene family
TPRKB	TP53RK binding protein	FGFR1	FGFR1	KIAA0101	p15(PAF)	RAN	RAN, member RAS oncogene family
TPRKB	TP53RK binding protein	MDM2	MDM2	TPX2	TPX2	RAN	RAN, member RAS oncogene family
TPRKB	TP53RK binding protein	PBK	PBK	XPO1	CRM1	RAN	RAN, member RAS oncogene family
TPRKB	TP53RK binding protein	TP53	p53	ZC3H11A	ZC3H11A	RAN	RAN, member RAS oncogene family
TPRKB	TP53RK binding protein	TPRKB	CGI-121	CENPA	CENP-A	SF3B1	splicing factor 3b subunit 1
TPRKB	TP53RK binding protein	ZFX	ZFX	EZH2	EZH2	SF3B1	splicing factor 3b subunit 1

Table 4.6: MetaCore Interactome results for the most significant interaction of the TCGA-STAD-MissenseTP53 cohort matching ANN driver analysis results

Input IDs	Network object name	Input IDs for correspondin g object name	Corresponding network object name	Input IDs	Network object name	Input IDs for correspondin g object name	Corresponding network object name
XPO1	exportin 1	BIRC5	Survivin	GSG2	GSG2	SF3B1	splicing factor 3b subunit 1
XPO1	exportin 1	CCND1	Cyclin D1	KIF11	KNSL1	SF3B1	splicing factor 3b subunit 1
XPO1	exportin 1	CDC25A	CDC25A	SETD1A	SET1A	SF3B1	splicing factor 3b subunit 1
XPO1	exportin 1	CHEK1	Chk1	ASPM	ASPM	TP73	tumor protein p73
XPO1	exportin 1	FGFR1	FGFR1	BAX	Bax	TP73	tumor protein p73
XPO1	exportin 1	MDM2	MDM2	BCL2	Bcl-2	TP73	tumor protein p73
XPO1	exportin 1	ORC1L	ORC1L	BCL2L1	Bcl-XL	TP73	tumor protein p73
XPO1	exportin 1	PTEN	PTEN	BUB1	BUB1	TP73	tumor protein p73
XPO1	exportin 1	RAD51	Rad51	CCNB1	Cyclin B1	TP73	tumor protein p73
XPO1	exportin 1	RAN	Ran	CDK1	CDK1 (p34)	TP73	tumor protein p73
XPO1	exportin 1	RANBP2	RanBP2	CDK4	CDK4	TP73	tumor protein p73
XPO1	exportin 1	TP53	p53	CDK6	CDK6	TP73	tumor protein p73
XPO1	exportin 1	TP73	p73	MAD2L1	MAD2a	TP73	tumor protein p73

Table 4.6: MetaCore Interactome results for the most significant interaction of the TCGA-STAD-MissenseTP53 cohort matching ANN driver analysis results

Input IDs	Network object name	Input IDs for correspondin g object name	Corresponding network object name	Input IDs	Network object name	Input IDs for correspondin g object name	Corresponding network object name
XPO1	exportin 1	ZFX	ZFX	MDM2	MDM2	TP73	tumor protein p73
ZC3H11 A	zinc finger CCCHtype containing 11A	CCNF	Cyclin F	RAD51	Rad51	TP73	tumor protein p73
ZC3H11 A	zinc finger CCCHtype containing 11A	CDK1	CDK1 (p34)	RCHY1	PIRH2	TP73	tumor protein p73
ZC3H11 A	zinc finger CCCHtype containing 11A	E2F1	E2F1	SASS6	SASS6	TP73	tumor protein p73
ZC3H11 A	zinc finger CCCHtype containing 11A	FANCD2	FANCD2	SERPINE1	PAI1	TP73	tumor protein p73
ZC3H11 A	zinc finger CCCHtype containing 11A	FGFR1	FGFR1	SFN	14-3-3 sigma	TP73	tumor protein p73
ZC3H11 A	zinc finger CCCHtype containing 11A	KIAA0101	p15(PAF)	STAG1	STAG1	TP73	tumor protein p73

Table 4.6: MetaCore Interactome results for the most significant interaction of the TCGA-STAD-MissenseTP53 cohort matching ANN driver analysis results

Input IDs	Network object name	Input IDs for correspondin g object name	Corresponding network object name	Input IDs	Network object name	Input IDs for correspondin g object name	Corresponding network object name
ZC3H11 A	zinc finger CCCHtype containing 11A	RAD54L	ATRX	TNFRSF10B	DR5(TNFRSF10B)	TP73	tumor protein p73
ZC3H11 A	zinc finger CCCHtype containing 11A	RAN	Ran	TP53	p53	TP73	tumor protein p73
ZC3H11 A	zinc finger CCCHtype containing 11A	SERPINB5	Maspin	TP53AIP1	P53AIP1	TP73	tumor protein p73
ZC3H11 A	zinc finger CCCHtype containing 11A	SGOL2	SGOL2	EZH2	EZH2	ZNF638	zinc finger protein 638
ZC3H11 A	zinc finger CCCHtype containing 11A	ZFX	ZFX	SF3B1	SF3B1	ZNF638	zinc finger protein 638

Orange colour indicates novel inputs and blue colour indicates known TP53 pathway members.

4.5.4. Combined (three-project) analysis for MissenseTP53 differential drivers

The differential source drivers associated with the TP53 pathway in the Missense TP53 status of the three projects (COADREAD-PAAD, and STAD) were combined to identify the common source drivers in the three cohorts (the TCGA-COADREAD-MissenseTP53, the TCGA-PAADMissenseTP53, and the TCAG-STAD-MissenseTP53). The commonalities were calculated using R programme (https://www.R-project.org/). The results revealed five sources that are novel to the TP53 pathway, three of which (SGOL2, TRAIP, and TACC3) are common between the TCGA-PAAD-MissenseTP53 and the TCGA-CRC-MissenseTP53, and two of which (CCDC150 and ERCC6L) are common between the TCGA CRC-MissenseTP53 and the TCGA-STAD-MissenseTP53.

A similar analysis was also done for the differential target drivers of the three projects. Three of the five common source drivers also appeared as common target drivers, which may reflect a homeostatic role of these drivers. Only one novel target driver (CHAF1B) appeared to be common between the TCGA-COADREAD-MissenseTP53 and the TCGA-STAD-

MissenseTP53 cohorts. The chromatin assembly factor 1, subunit B (CHAF1B), plays an important role in chromatin assembly and DNA replication during proliferation. This protein is also involved in DNA repair, and it has been previously linked to cancer in previous research. A literature search for the terms “CHAF1B” and “cancer” revealed 17 publications, one of which revealed a prognostic ability of CHAF1B in gastric cancer (Ren et al., 2022). Another study suggested CHAF1B among the genes that have a unique association between their mRNA expression and their knockdown/ knockout efficacy in colon cancer cell lines (Jeong et al., 2020). However, none of the previous publications relate CHAF1B to the Missense TP53 mutation status in cancer, which indicates that this is a novel finding of the ANN data mining tools.

The CHAF1B combined subnetwork is presented as a Cytoscape image in Figure 4.16. Further downstream analysis for more confirmation was also done using the Human Atlas Database to identify CHAF1B protein expression in colorectal and stomach cancers. The protein is highly expressed in 12 out of 12 colorectal cancer cases, and in 10 out of 12 stomach cancer cases. Figures 4.17 and 4.18 presents immune-stained stained slides for colorectal and stomach cancers, respectively, adapted from the Human Protein Atlas database (The Human Protein Atlas). Known TP53 pathway members also appeared among multiple cohorts, including FAS and CDK2, as common source drivers and CDKN2A as a common target driver.

The results of the combined analysis are presented in Tables 4.7 and 4.8.

Table 4.7: Combined driver analysis for the three cohorts (Sources)

Influencer (Sources)					
Gene Symbol	**Gene Name**	**Project**	**Rank**	**Sum of average**	**ABS**
SGOL2	Shugoshin 2	COADREAD	8	-0.3777	0.3777
		PAAD	24	35.6884	35.6884
ERCC6L	ERCC Excision Repair 6 Like, Spindle Assembly Checkpoint Helicase	COADREAD	30	-21.9703	21.9703
		PAAD	10	7.8956	7.8956
TACC3	Transforming Acidic Coiled-Coil Containing Protein 3	COADREAD	34	-23.3711	23.3711
		PAAD	41	10.5254	10.5254
CCDC150	Coiled-Coil Domain Containing 150	COADREAD	47	-31.7806	31.7806
		STAD	28	6.2453	6.2453
TRAIP	TRAF Interacting Protein	COADREAD	52	-35.7583	35.7583
		PAAD	26	33.4327	33.4327
FAS	Fas Cell Surface Death Receptor	COADREAD	40	-28.5661	28.5661
		STAD	25	6.3965	6.3965
CDK2	Cyclin Dependent Kinase 2	PAAD	39	17.3323	17.3323
		STAD	22	6.5087	6.5087

Orange colour indicates novel sources and the blue colour indicates known pathway members

Table 4.8: Combined driver analysis for the three cohorts (Targets)

Influenced by (Targets)					
Gene Symbol	**Gene Name**	**Cohort**	**Rank**	**Sum of average**	**ABS**
TACC3	Transforming Acidic Coiled-Coil Containing Protein 3	COADREAD	19	-11.184	11.2
		PAAD	66	-40.204	40.2
SGOL2	Shugoshin 2	COADREAD	68	-51.635	51.6
		PAAD	40	14.124	14.1
TRAIP	TRAF Interacting Protein	COADREAD	79	-61.702	61.7
		PAAD	51	-5.461	5.46
CHAF1B	Chromatin Assembly Factor 1 Subunit B	COADREAD	78	-59.220	59.2
		STAD	18	-1.431	1.43
CDKN2A	Cyclin Dependent Kinase Inhibitor 2A	COADREAD	29	-20.44	
		PAAD	9	102.85	
		STAD	1	65.51	

Orange colour indicates novel sources and the blue colour indicates known pathway members

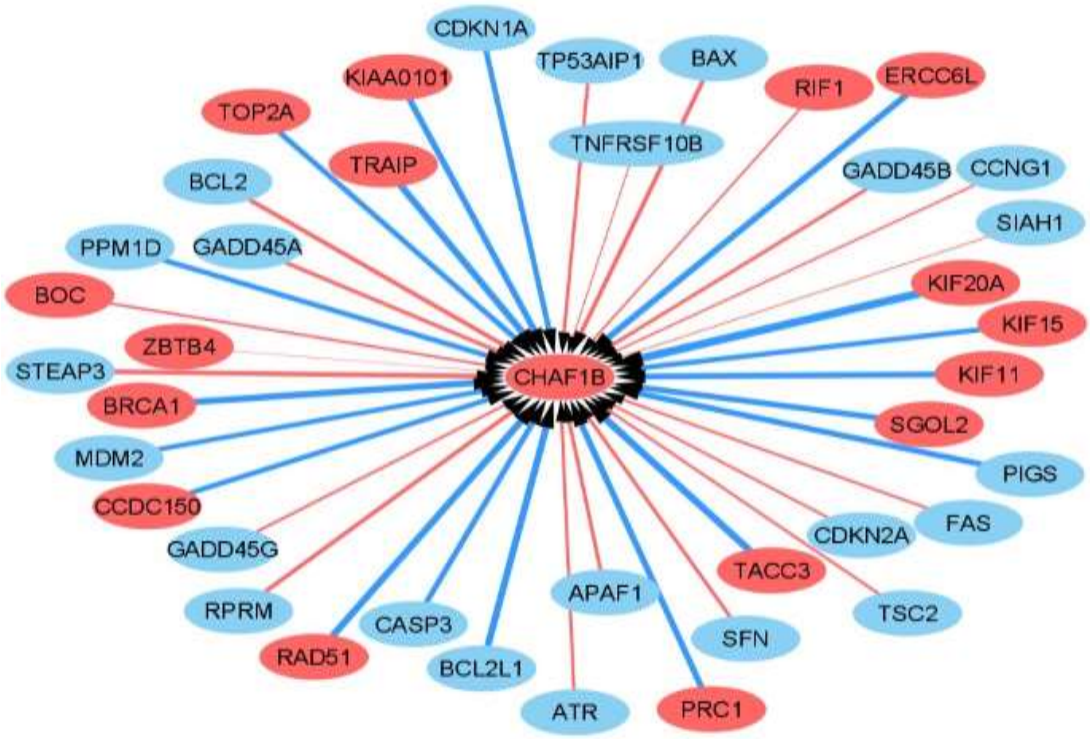
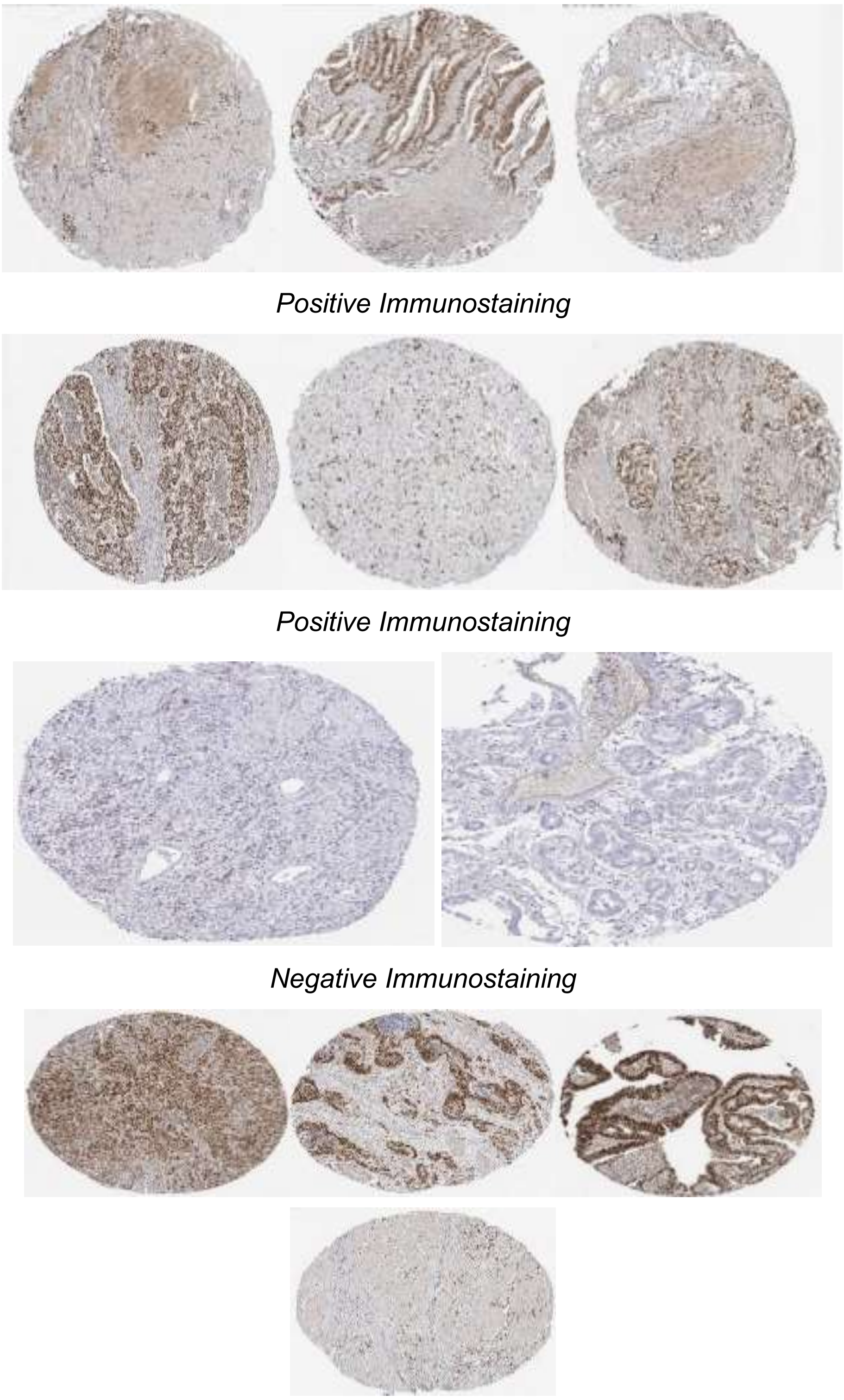


Figure 4.17: CHAF1B combined subnetwork for the COADREAD and STAD cohorts

Shows CHAF1B as a hub target, blue colour marks for known TP53 members, red colour for genes not belonging to the pathway, blue edge for positive interactions, and red for negative interactions. The figure shows 18 negative and 6 positive interactions with known pathway targets, and 3 negative and 13 positive negative interactions with non-pathway members



Positive Immunostaining

Positive Immunostaining

Negative Immunostaining

Positive Immunostaining

Figure 4.18: Human Protein Atlas results for CHAF1B protein expression shows strong immunostaining in 10 out of 12 examined stomach cancer cases

Source: adapted from Human Protein Atlas ([Expression of CHAF1B in stomach cancer - The Human Protein Atlas])

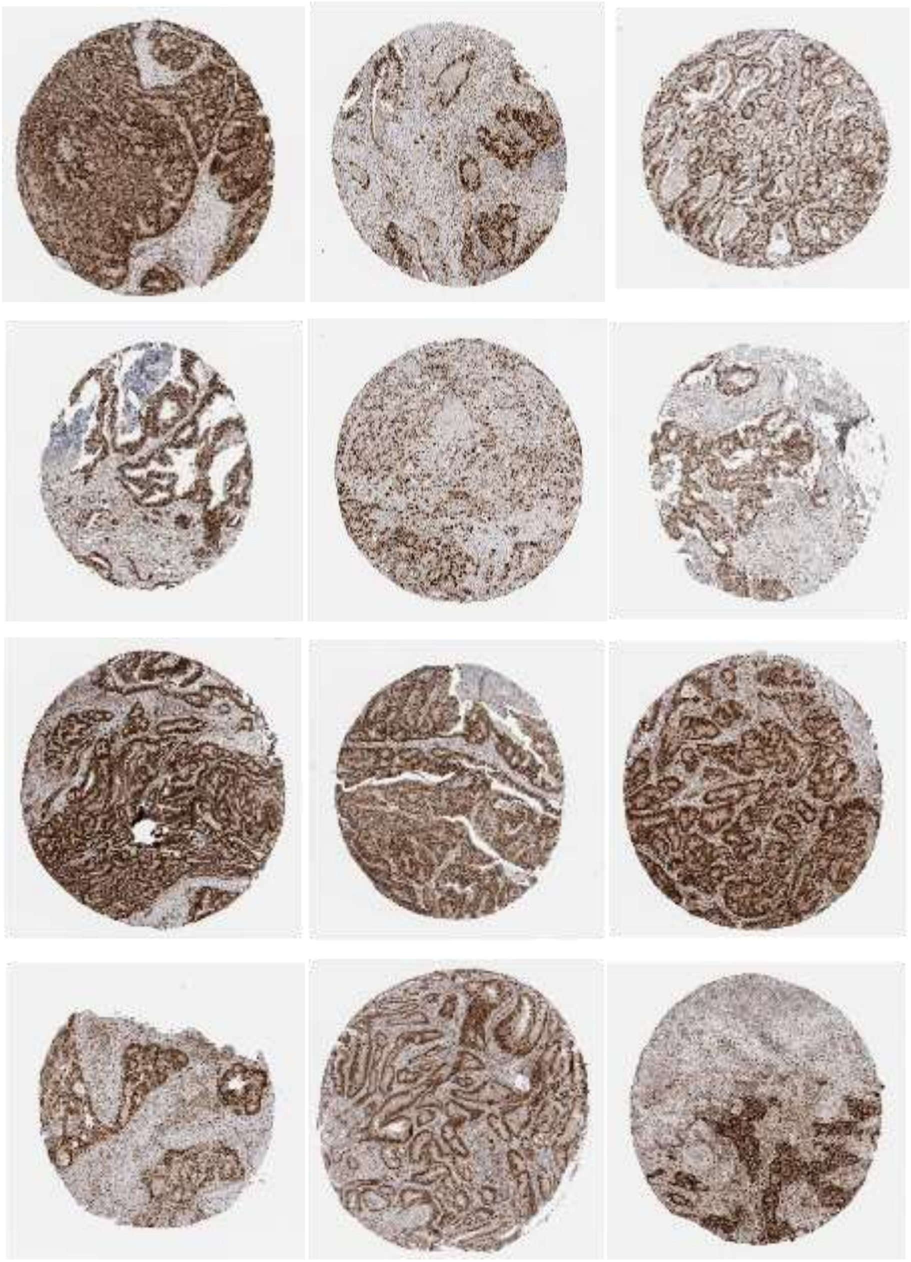


Figure 4.19: Human Protein Atlas results for CHAF1B protein expression shows strong immunostaining in 12 out of 12 examined colorectal cancer cases

Source: adapted from Human Protein Atlas ([Expression of CHAF1B in colorectal cancer - The Human Protein Atlas]())

4.6. Summary and conclusion

In this chapter, ANN stepwise algorithm was used for the analysis of the TP53 pathway based on the mutation status of the TP53 gene. Data were obtained from three projects of the TCGA data (COADREAD, PAAD, and STAD). Two cohorts (Mutant- and Wild-type TP53) were identified and analysed for each project separately. The analysis spanned 64 known TP53 pathway members, each of which was considered as a separate ANN model, to identify genes that were most related to that member. The top-200 commonalities for all members were identified, then a comparative analysis was undertaken to identify common and distinctive predictors with significant differential expression between the Mutant- and the Wild-type TP53 cohorts.

The results showed 65 distinctive predictors associated with the TCGA-COADREAD Mutant TP53, 100 for the TCGA-PAAD Mutant TP53 cohorts, and 241 for the TCGA-STAD Mutant TP53 cohort. The distinctive predictors were then used to build the network of interaction using the ANNI algorithm, and driver analysis was applied to identify the differential drivers associated with the pathway in the MissenseTP53 mutation status for each project. Interactome analysis was performed using MetaCore platform. ANN driver results were then compared to MetaCore Interactome results for each project to identify the concordance between the two methods.

4.6.1. TCGA-COADREAD-MissenseTP53 cohort

4 genes that appeared as top differential source drivers also appeared as significant network objects (DUSP4, PHLDA1, KIF20A, TOP2A) in the MetaCore Interactome result. A similar number of top differential target drivers (GIFI1, CASP3, MAP3K3, DDB2) were also found as corresponding significant objects in MetaCore interaction results. A literature search suggest the these genes have been linked to Colorectal cancer except GIFI1 which has been known as a regulator for myeloid cell differentiation and proliferation and has been correlated with favourable prognosis in Acute myeloid leukaemia (Salarpour et al., 2020) however, no previous report linked GIFI1 with Colorectal cancer .

4.6.2. TCGA-PAAD-MissenseTP53 cohort

The MetaCore result indicated 9 significant network objects which also appeared among the top differential sources in ANN driver analysis result (C15orf42, CENPN, E2F7, FANCD2, GTSE1, MCM10, MCM2, MCM4 and SPAG5), and 8 significant corresponding

MetaCore objects appeared among the top target drivers in the ANN driver analysis results (CDK1, CDK2, CDK6, CCMB1, FANCA, GADD45A, RRM2B, TANK).

4.6.3. TCGA-STAD-MissenseTP53 cohort

MetaCore Interactome analysis indicated 13 significant network objects which were also appeared among the top differential sources in the ANN driver analysis results (CAND1, CCNE1, CCNT2, ERCC6L, FOSL1, FOXN2, GADD45A, NUP107, RFC5, RFWD2, TPRKB, XPO1, ZC3H11A), and 12 significant corresponding MetaCore objects also appeared among the top differential targets in ANN driver analysis (APAF1, CDKN2A, GORAB, LRDD, MDM2, PBK, PMAIP1, RAN, SF3B1, TP73, PPM1D, ZNF638).

4.6.4. Combined analysis

Combined analysis for the differential drivers of the three projects was done to identify common drivers between all sets. The results revealed two sources (CCDC150 and ERCC6L) that are common between the TCGA-COADREAD-MissenseTP53 and the TCGA-STADMissenseTP53 and are novel to the TP53 pathway. **A novel target driver (CHAF1B)** appeared to be common between the TCGA-COADREAD-MissenseTP53 and the TCGA-

STAD-MissenseTP53 cohorts. The CHAF1B protein is highly expressed in colorectal and stomach cancers based on the Human Protein Atlas results. Although it has been linked previously to poor outcome in gastric cancer, no previous association between CHAF1 and missense TP53 mutation status.

Overall, this chapter provides more evidence for the practicality of using ANN data mining tools for pathway modelling. The concordance between the ANN and MetaCore results provides an extra layer of strength and gives more indication about the reliability of ANN results. The identification of significant predictors that are highly associated with the TP53 pathway in the mutant and the wild-type state of the TP53 gene also represents a strong point of the study since it shows that the pathway may be activated differently based on the TP53 mutation, the extension of the analysis to model the interaction and to recognise the potential drivers related to the pathway by considering the missense mutation status of the TP53 gene for each of the investigated projects also reflect a further benefit of the research. The combined analysis revealed some similarity between the examined cohorts and that could help in better understanding of the behaviour of the pathway in the missense mutation status of the TP53 gene. However, more similarity might be yielded by considering more data related to other cancer types, this represent an area for improvement of the analysis. Also, the results were only focused on the Missense TP53 mutation status. They did not consider other mutation types since they have low sample numbers, which cannot be analysed using the ANNI algorithm. Chapter 5 seeks to build an interaction network and driver analysis for the TP53 pathway in the Wild-type TP53 status in the three examined cohorts.

CHAPTER 55 ANN INFERENCE MODELLING INTERACTION AND IDENTIFICATION OF TP53 PATHWAY DIFFERENTIAL DRIVERS IN WILD-TYPE

5.1. Introduction

The previous chapter presented the application of ANN approaches for modelling the TP53 pathway based on the mutation status of the TP53 gene using data from three TCGA projects (COADREAD, PAAD, and STAD). The results indicated a panel of distinctive predictors associated with the TP53 pathway in Mutant- and Wild-type mutation status of the TP53 gene.

The interaction network leads to the identification of differential drivers associated with the TP53 pathway in the MissenseTP53 mutation status for each project separately. The combined analysis revealed common drivers associated with the pathway in the MissenseTP53 state for different projects. This chapter models the interaction for the distinctive predictors associated with the TP53 pathway in the Wild-type state of the TP53 for the three TCGA projects and identifies the concordant drivers between all projects.

The Wild-type TP53 has important physiological roles within the cell. It becomes activated upon cellular stress, causing the cell cycle to stop repair or eventual cell death if the damage is severe. This process is facilitated by various transcriptional factors that interact with the Wild-type TP53 to activate or repress appropriate downstream target genes in cancer. The Wild-type TP53 induces the expression of genes that inhibits cancer growth and progression. It is likely that abnormal signalling of the TP53 pathway occurs in cancer that carries Wild-type TP53. In recent years, studies in the field have supported the development of new therapeutic approaches that target Mutant- and Wild-type TP53 to suppress tumour growth (Babamohamadi et al., 2022). Consequently, it is crucial to model the interaction associated with the pathway in the Wild-type TP53 status.

5.2. Chapter aims

This chapter is an extension of the analysis that has been done in chapter 4. It builds interaction networks and identifies molecular drivers using the differential predictors associated with the TP53 pathway in the Wild type state of the TP53 genes for the three TCGA projects (COADREAD, PAAD and STAD). These predictors have been Identified and mentioned in sections 4.6.1, 4.6.2, and 4.6.3. Moreover, this chapter supports the project's overall goal by demonstrating that ANN-based data mining techniques could be used as a tool for pathway modelling. This chapter is discrete from Chapter 4 since it is focused on modelling of interaction associated with the TP53 pathway in its wild-type state. This is crucial for understanding normal cellular processes, discovering therapeutic targets, identifying dysregulation in cancer and gaining insights into cancer biology. By studying TP53 interaction network in the wild-state, it is possible to distinguish between normal pathway regulation and aberrant signalling caused by mutations. This helps in identifying specific alterations associated with cancer development.

5.3. Chapter objectives

- Build ANN of interaction (ANNI) for the Wild-type TP53 cohort for each project.
- Predict the key interactions associated with the pathway in the Wild type TP53 state for each project.
- Identification of differential drivers connected to the pathway in the Wild type TP53 state for each project.
- Combined analysis to identify similarity in drivers between different projects.

5.4. Methods

This analysis involves building of the interaction network for the distinctive predictors associated with the wild type TP53 which obtained through the application of the Stepwise ANN approach in section 4.6.1. ANN network inference and driver approaches were applied following the protocols described in chapter3. Figure 5.1 provides a schematic representation for the analysis stages.

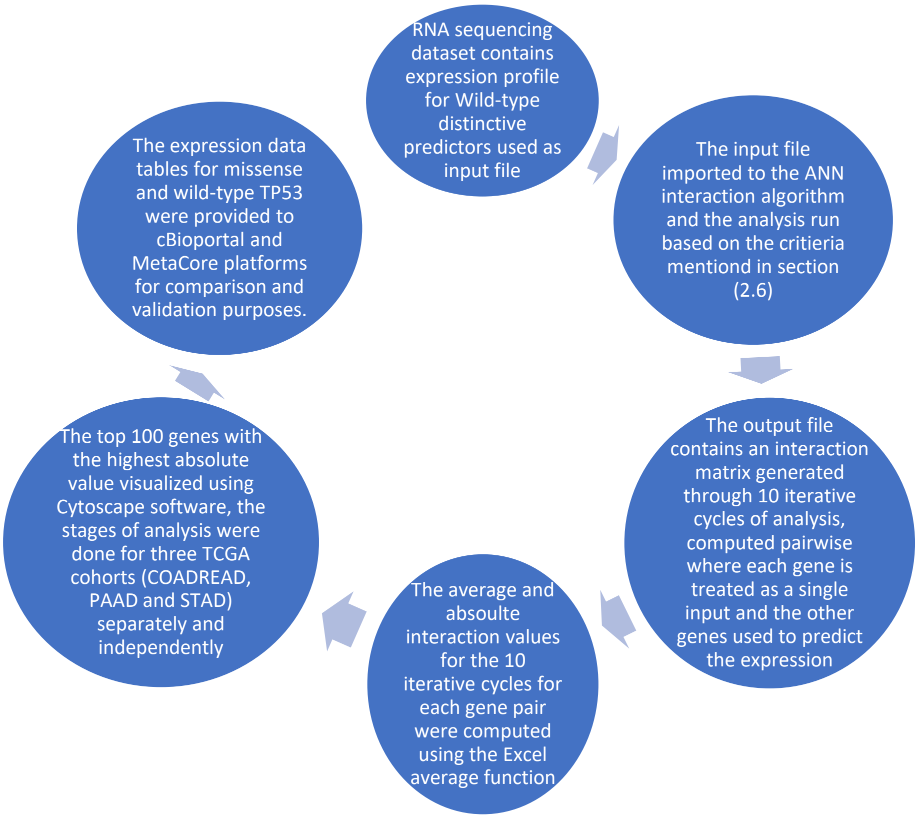

RNA sequencing dataset contains expression profile for Wild-type distinctive predictors used as input file
The input file imported to the ANN interaction algorithm and the analysis run based on the critieria mentiond in section (2.6)
The output file contains an interaction matrix generated through 10 iterative cycles of analysis, computed pairwise where each gene is treated as a single input and the other genes used to predict the expression
The average and absoulte interaction values for the 10 iterative cycles for each gene pair were computed using the Excel average function
The top 100 genes with the highest absolute value visualized using Cytoscape software, the stages of analysis were done for three TCGA cohorts (COADREAD, PAAD and STAD) separately and independently
The expression data tables for missense and wild-type TP53 were provided to cBioportal and MetaCore platforms for comparison and validation purposes.

Figure 5.1: A schematic representation for the stage of analysis.

5.5. Results and Discussion

5.5.1. TCGA-COADREAD-WTTP53 cohort

The distinctive predictors associated with the TP53 pathway for the TCGA-COADREAD Wildtype TP53 cohort were identified in the previous chapter. These predictors were used to build the network of interaction using the ANNI algorithm, and a total of 107 inputs were used to run the interaction analysis following the protocol described previously, which produced a matrix of (107x(107-1))11130 interactions; the results were filtered, and drivers were identified using the method described in section 3.5.3. A rank order was devised based on the highest sum of interaction values. The top-100 interactions indicated seven hub targets, four of which are known pathway members (GADD45B, ZMAT3, BAS, and SIAH1), and three of which (RPS27L, SNRPD1, and MPDU1) are novel to the pathway.

The results of the top-100 interactions were presented using Cytoscape, as shown in Figure 5.2. RPS27L appeared as a major subnetwork with 10 genes not belonging to the pathway and one interaction with a known pathway member (BAX), which appeared as a hub node. There were overlaps between the two subnetworks in three genes (UBE2N, BOLA3, and FASTKD1). RAPS27L was identified among the genes that are common between five colorectal datasets in the analysis presented in section 3.6.2. RPS27L protein was more prevalently expressed in the WTTP53 compared to the Missense TP53 cohort, according to the cBioPortal result. The association between RPS27L and Bax (a well-known apoptosis regulator factor) in the presented ANN result was related to He and Sun's (2007) finding that RPS27L was among the ribosomal proteins that regulate the TP53 function and a direct TP53 target that mediates p53-induced apoptosis. Moreover, differential drivers associated with the pathway in the TCGA-COADREAD-WTTP53 cohort are identified and presented in Figures 5.5 and 5.6. The differential drivers include members of the TP53 pathway found to be top targets (GADD45A and BAX); and members who showed as top sources (ATM and BCL2). Novel genes were also found, including the top target drivers (TRIAP1 and CDCA2), and the top source drivers (ATRX and ZNF445). RAPS27L also appeared as a top target driver that is negatively influenced by other genes in the network.

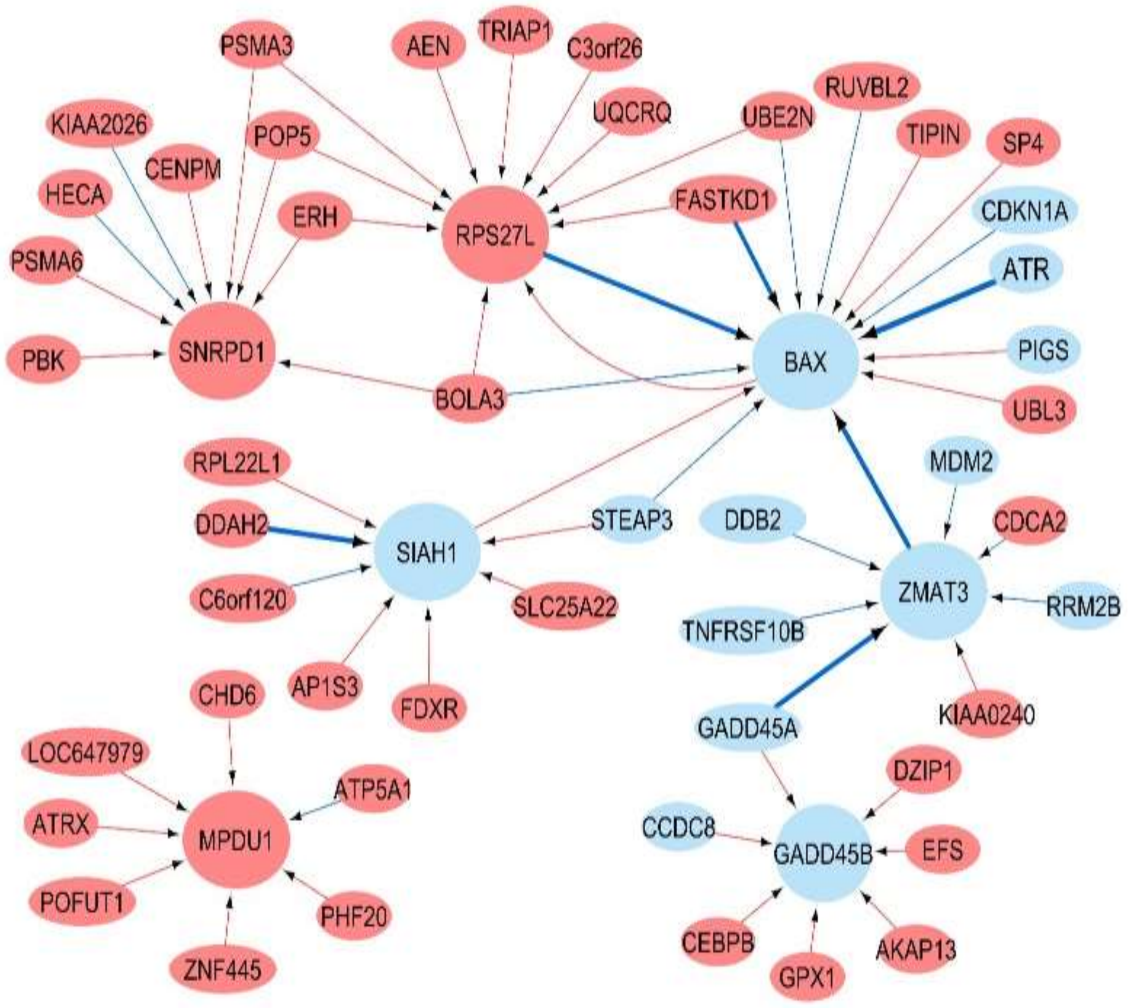


Figure 5.2: Cytoscape image for the TCGA-COADREAD-WTTP53 cohort

Blue nodes indicate known TP53 targets, and the red nodes indicate new candidates. The blue lines indicate positive interactions, and the red lines indicate negative interactions. Line thickness indicates interaction strength. Hub nodes are those with more than 5 interactions

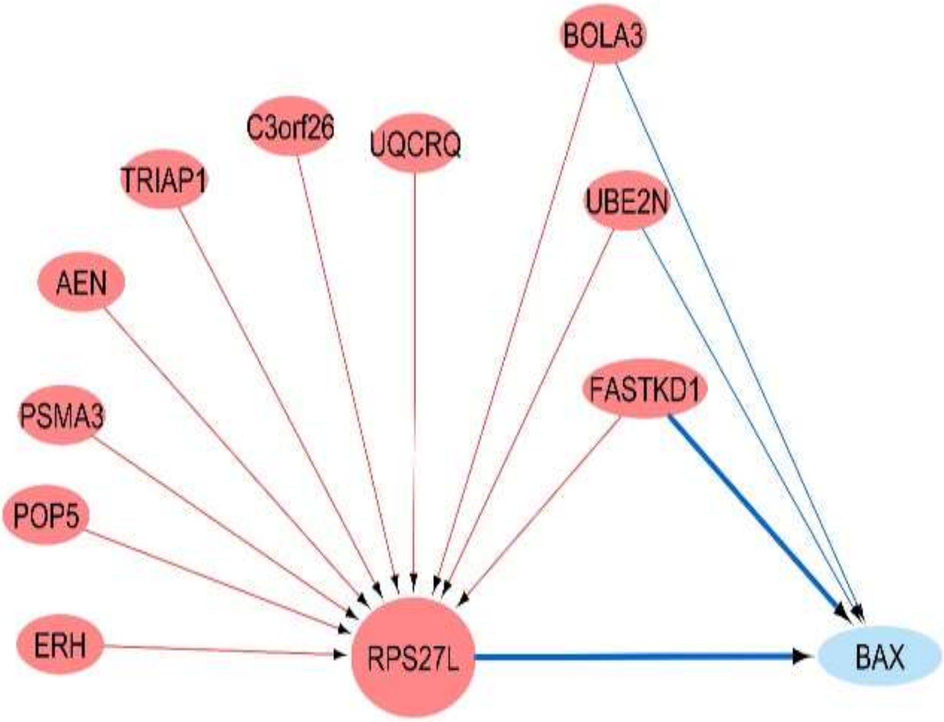


Figure 5.3: RPS27L subnetwork

Represents the major novel hub node with 10 new interactions with non-pathway members (shown as red nodes) and one interaction with known pathway member (BAX) (shown as blue node)

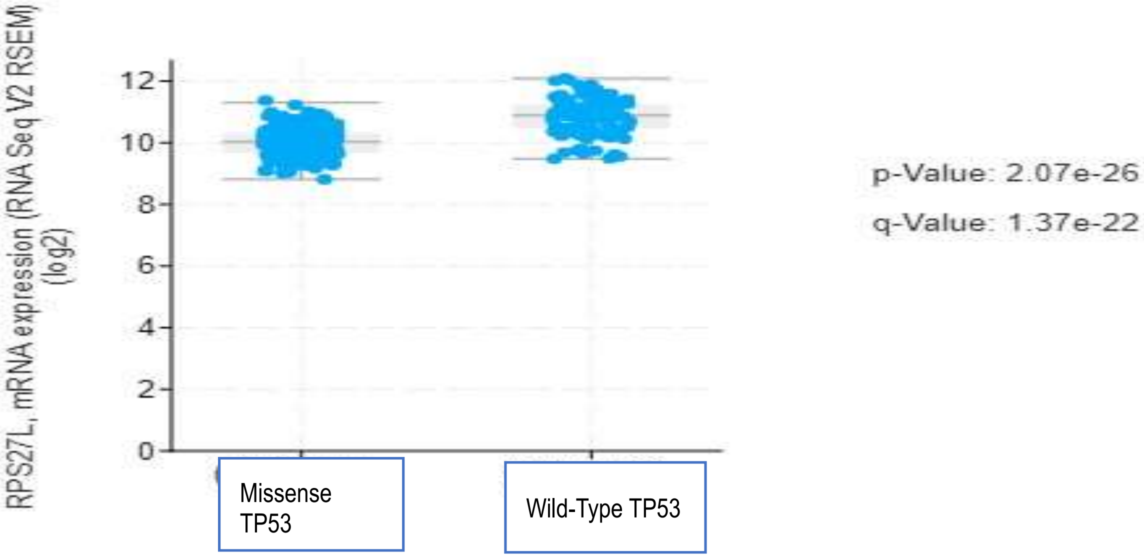


Figure 5.4: Expression of RPS27L gene in cBioPortal

The gene is more highly expressed in the TCGA-COADREAD WTTP53 cohort compared to the TCGA-COARDEAD-MissenseTP53 cohort Source: cBioPortal

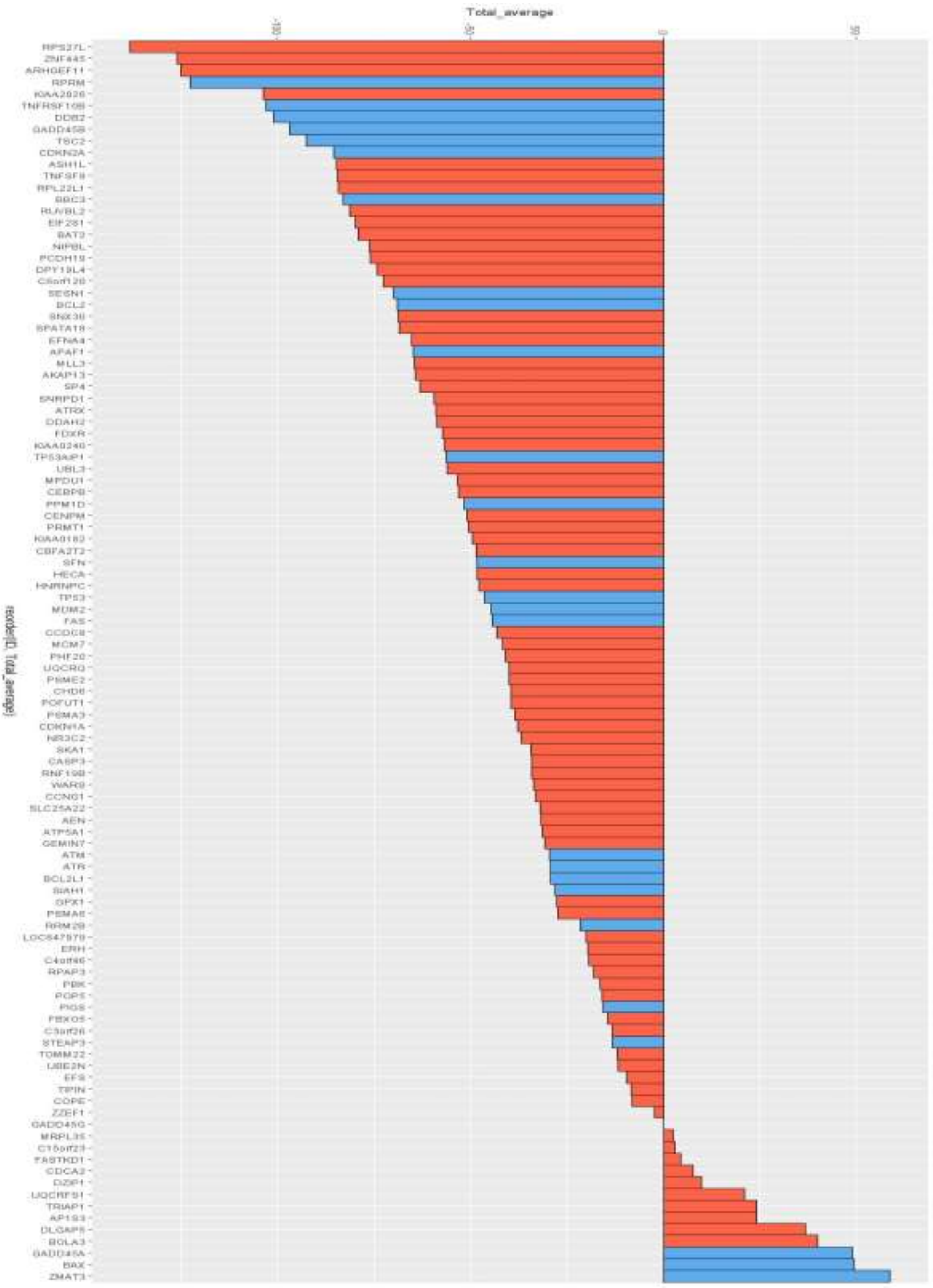


Figure 5.5: Differential drivers (as targets) associated with the TCGA-COADREAD-WTTP53

Sorted based on the total average of interactions. Blue colour indicates known TP53 pathway members, and red colour indicates novel targets

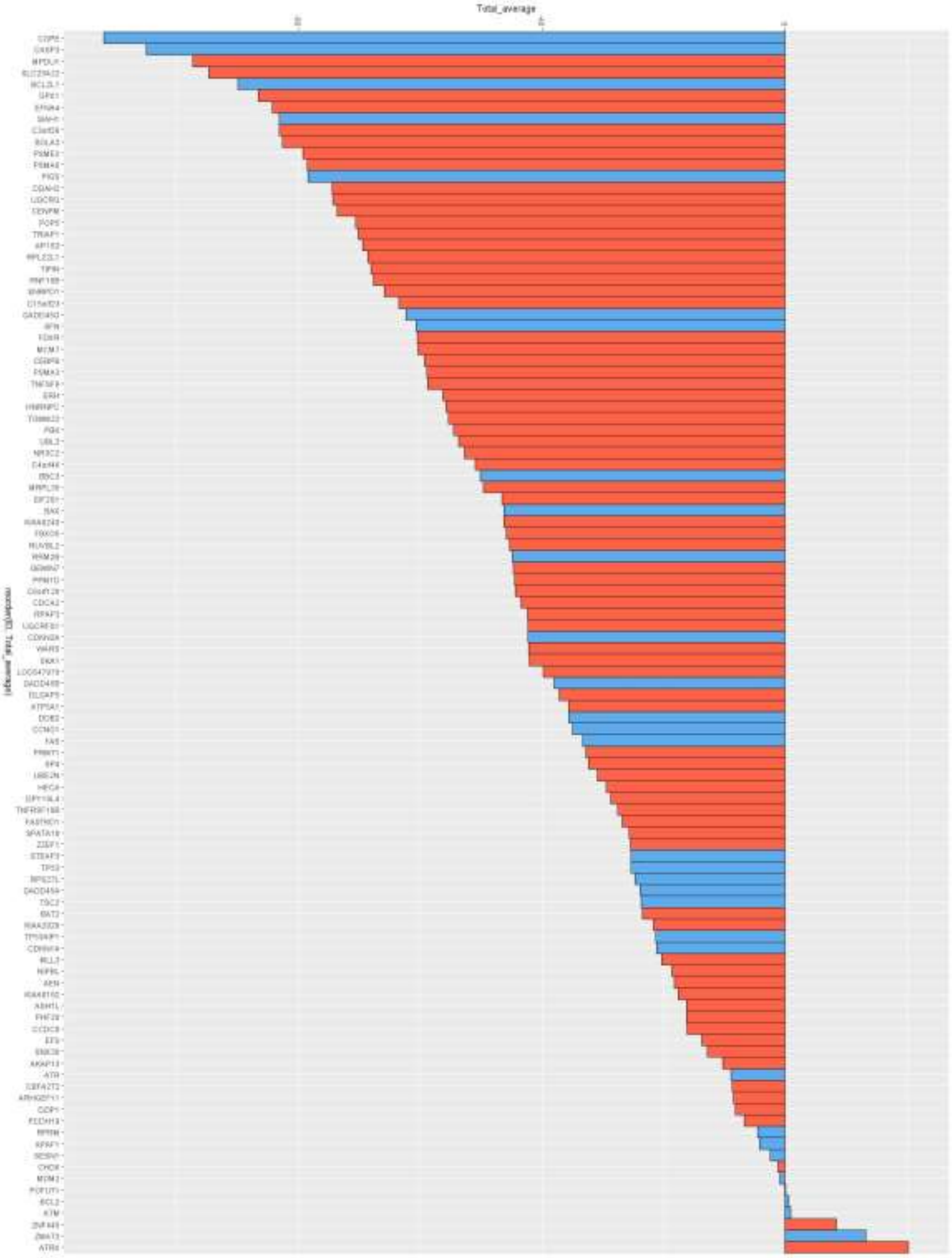


Figure 5.6: Differential drivers (as sources) associated with the TCGA-COADREAD-WTTP53

Sorted based on the total average of interactions. Blue colour indicates known TP53 pathway members, and red colour indicates novel targets

5.5.2. TCGA-PAAD-WTTP53 cohort

The analysis of the TCGA-PAAD-WTTP53 cohort was done using the distinctive predictors identified in section 4.5. The interaction network was done using the ANNI protocol explained in section 3.4.2, producing a matrix of (195 x (195-1)) 37830 interactions. The results were filtered out based on the highest sum of interactions. NXF2B, CDCA2, FM01, and TMPRSS4 were found to be the major hub nodes for the top-100 interactions. NXF2B represented the subnetwork that was highly influenced by other genes, five of which were known pathway members, while 34 were new to the pathway. The cBioPortal result showed higher expression of NXF2B protein in the TCGA-PAAD-WTTP53 than in the MissenseTP53 cohort, as shown in Figure 5.7. NXF2B gene is related to progesterone-receptor negative breast cancer, based on the Human Disease Database search (Progesterone-Receptor Negative Breast Cancer disease: Malacards - Research Articles, Drugs, Genes, Clinical Trials).

No previous report was found on the association of NXF2B and pancreatic cancer, which could be a novel finding of the ANN inference approach. However, NXF2B does not appear among the top targets identified from the driver analysis results. By contrast, TMPRSS4 gene, which was found among the hub nodes for the top-100 interactions, was also found as a target differential driver, with the highest sum of the average interactions for the whole network A literature search also supports this finding; a remarkable overexpression of TMPRSS4 has been documented in pancreatic cancer tissue, and it plays a control role in cellular proliferation and apoptosis (Gu et al., 2021), indicating an important role in pancreatic cancer. Moreover, most of the identified top differential drivers did not belong to the TP53 pathway, including RHBDL2, MST1R, and IGFBP2 as the top targets; and NDE1, FAM83A, and OSEPL3 as the top sources. Figures 5.10 and 5.11 present the differential driver analysis results.

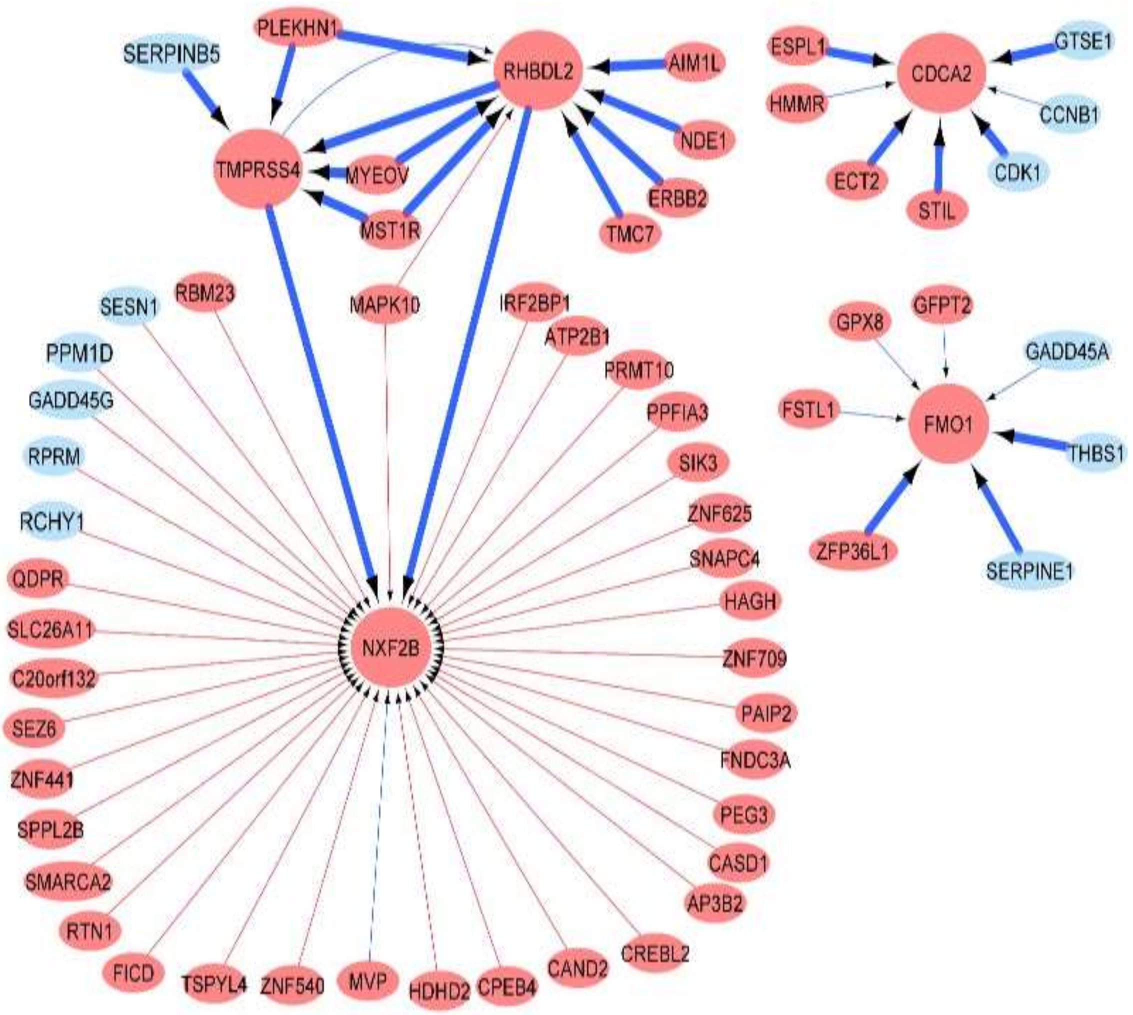


Figure 5.7: Cytoscape image for the TCGA-PAAD WTTP53 cohort

Blue nodes indicate known TP53 targets, and the red nodes indicate new candidates. Blue lines for positive interactions and the red lines indicate negative interactions. Line thickness indicates interaction strength. Hub nodes are those with more than 5 interactions

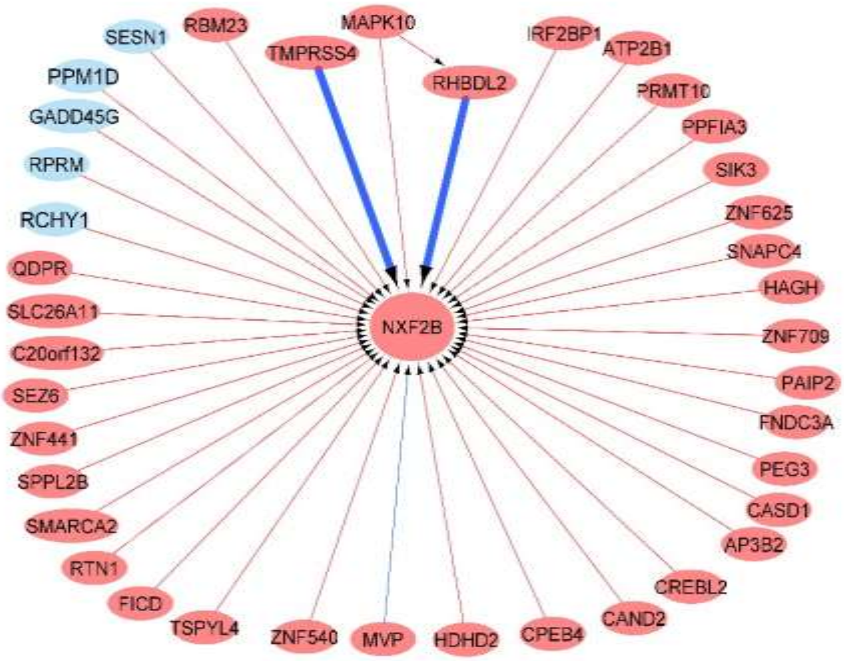


Figure 5.8: NXF2B subnetwork

Represents the major hub node, with 34 new interactions. Non-members of the pathway shown as red nodes, and 5 interactions with known pathway members shown as blue nodes.

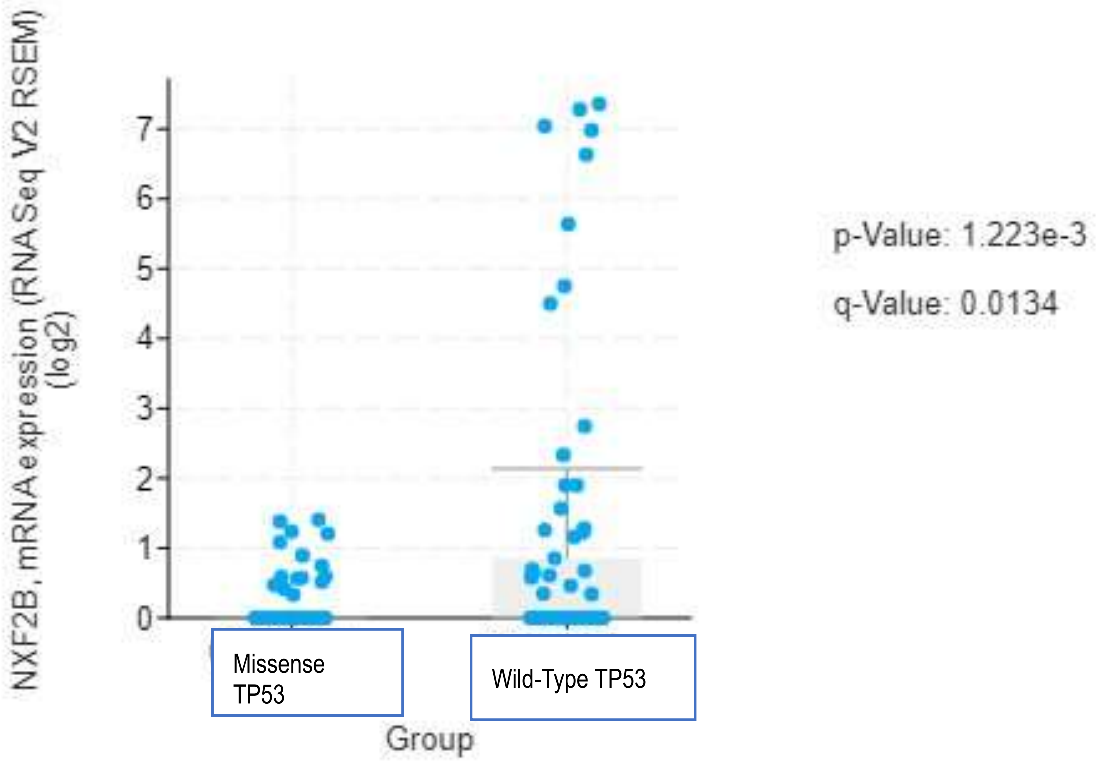


Figure 5.9: Expression of NXF2B gene in cBioPortal

The gene is more highly expressed in the TCGA-PAAD WTTP53 cohort compared to the TCGA-PAAD-MissenseTP53 cohort

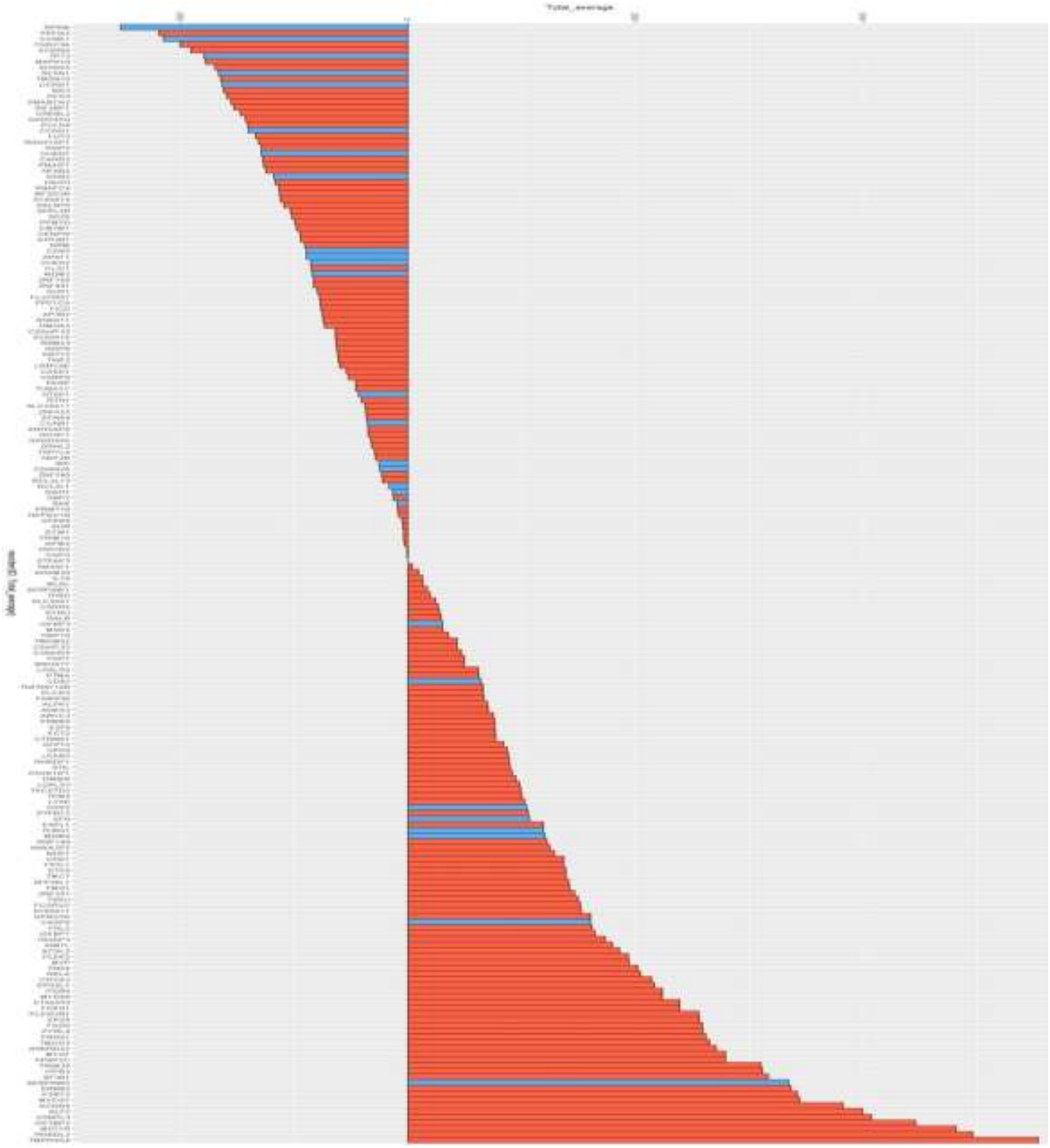


Figure 5.10: Differential drivers (as targets) associated with the TCGA-PAAD-WTTP53 Sorted based on the total average of interactions. Blue colour indicates known TP53 pathway members, and red colour indicates novel targets

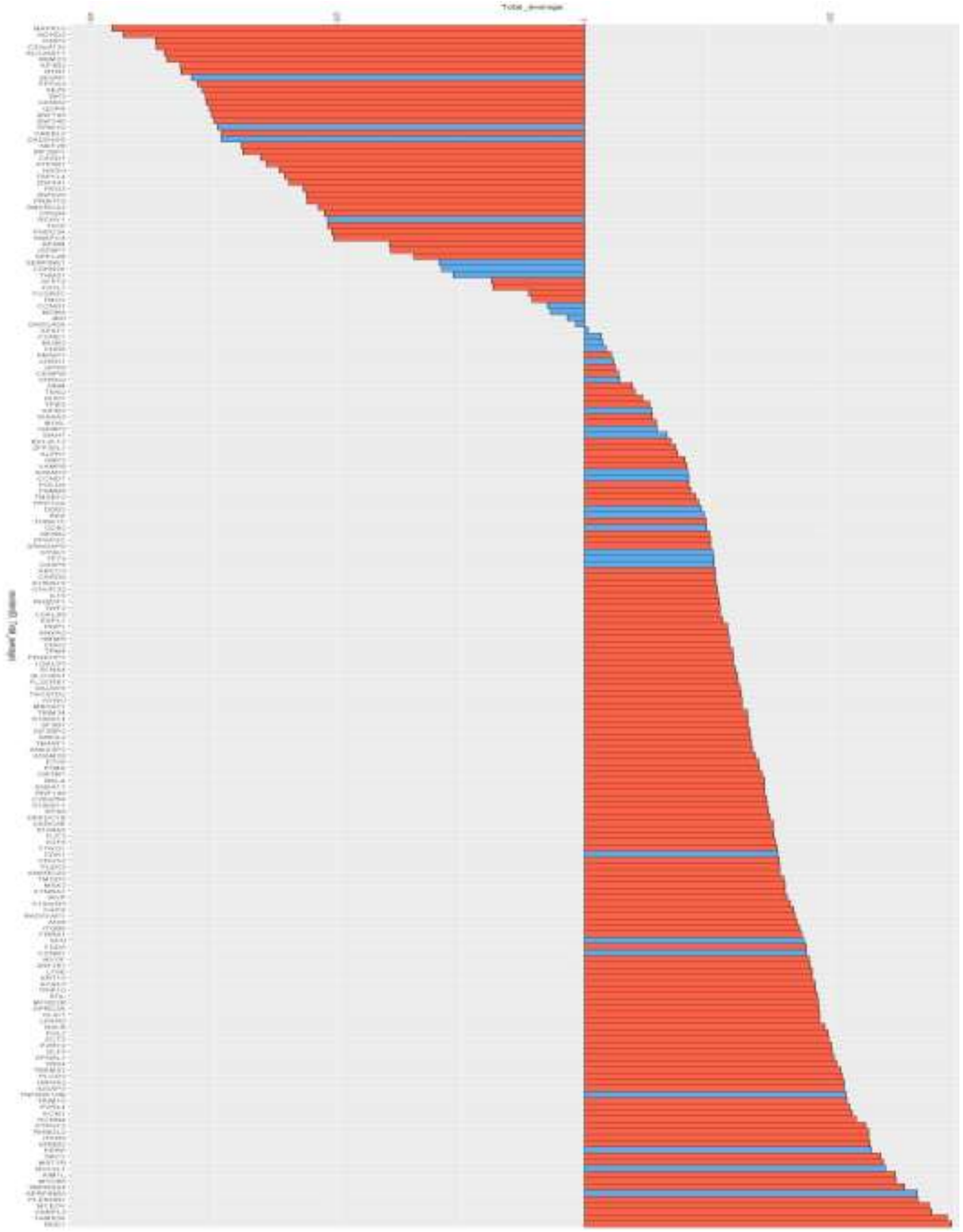


Figure 5.11: Differential drivers (as sources) associated with the TCGA-PAAD-WTTP53

Sorted based on the total average of interactions. Blue colour indicates known TP53 pathway members, and red colour indicates novel targets

5.5.3. TCGA-STAD-WTTP53 cohort

This analysis uses the distinctive predictors associated with the TP53 pathway for the TCGASTAD-WTTP53 cohort identified in section 4.5. The interaction network was built using ANNI approach following the protocol explained in section 3.4.2, creating a matrix of (422 x(4221))3569 interactions. These were filtered out based on the sum of the interactions. The results of the top-100 interactions revealed five main hub targets which do not belong to the TP53 pathway (FUT8, SNORD116-26, TSNAXIP1, CHSY1, and TRAM1L1). These targets have negative interactions, mainly with non-pathway members.

FUT8 represents the main hub node, with 35 interactions, the interesting observation about this subnetwork is that all the interactions are not members of the TP53 pathway. Which may indicate a non-direct relation of this gene with the TP53 pathway. FUT8 protein expression is significantly higher in the Wild-type TP53 compared to the Missense TP53 cohort, according to the cBioPortal results, as indicated in Figure 5.12. Upregulation of FUT8 has been described in different cancers, including gastric cancer, indicating a possible function in the regulation of tumour development and progression (Liao et al., 2021).

The driver analysis indicates novel differential elements that mostly appeared as sources that did not belong to the TP53 pathway. Among them, ARHGAP28, BTBD7, and AIM1L were the top differential targets; and CAPSL, C8orf48, and SNORD116.25 were the top differential sources. Known pathway members were also found mainly as top differential targets, including TSC2, BCL2L1, and TP73. The differential drivers are presented in Figures 5.15 and 5.16.

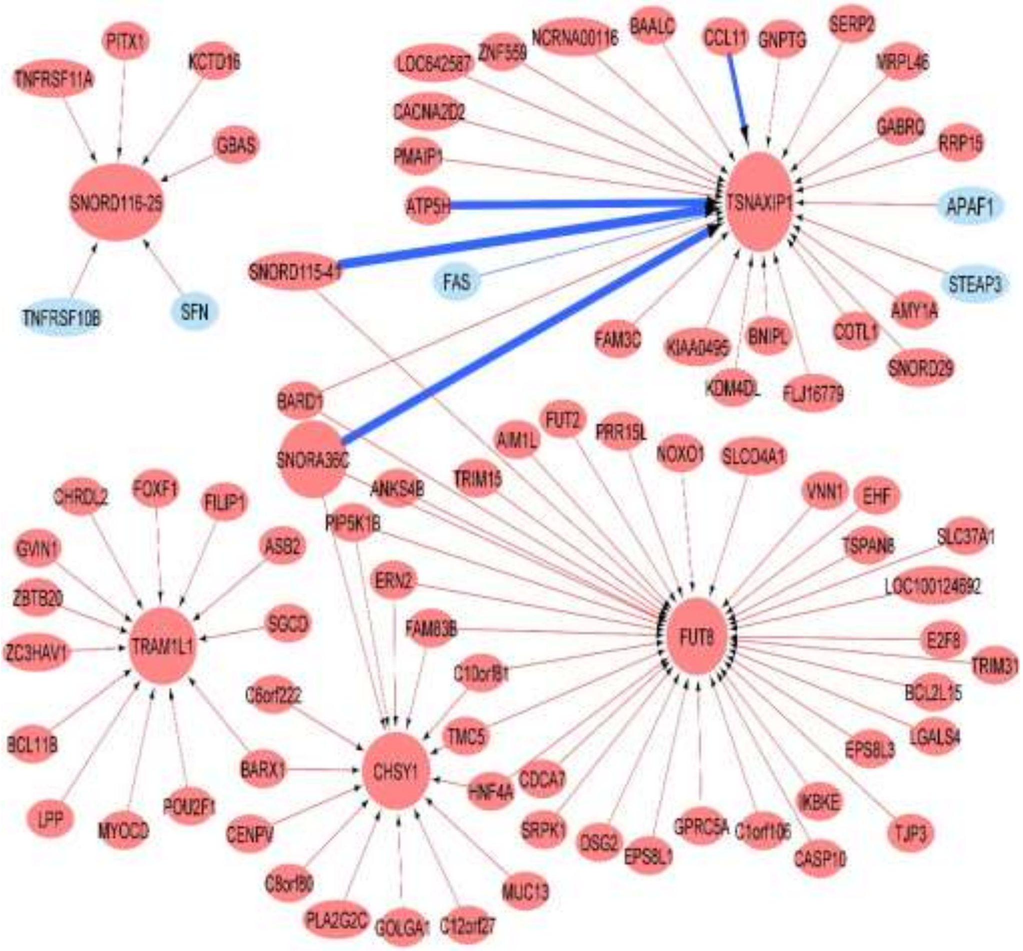


Figure 5.12: Cytoscape image for the TCGA-STAD WTTP53 cohort

Blue nodes indicate known TP53 targets while red nodes indicate new candidates. Blue lines indicate positive interactions, and red lines indicate negative interactions. Line thickness indicates interaction strength. Hub nodes are those with more than 5 interactions

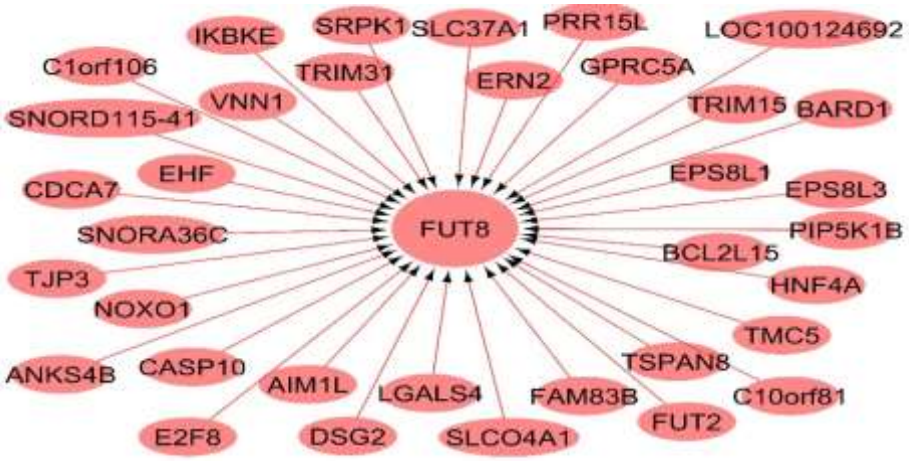


Figure 5.13: FUT8 subnetwork

Represents the major hub node with 35 new interactions with non-TP53 pathway members shown as red nodes

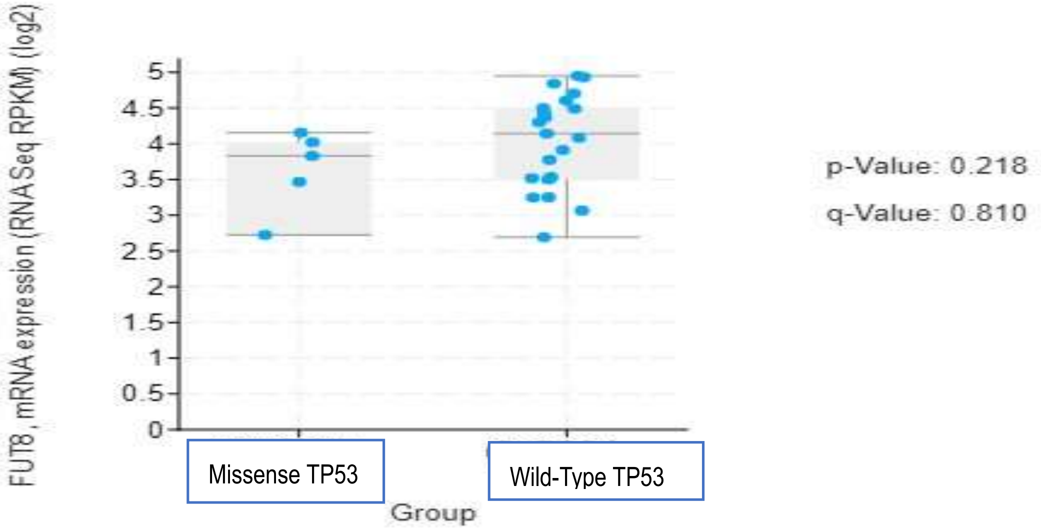


Figure 5.14: Expression of FUT8 gene in cBioPortal

The gene is more highly expressed in the TCGA-STAD WTTP53 cohort compared to the TCGA-STAD-MissenseTP53 cohort.

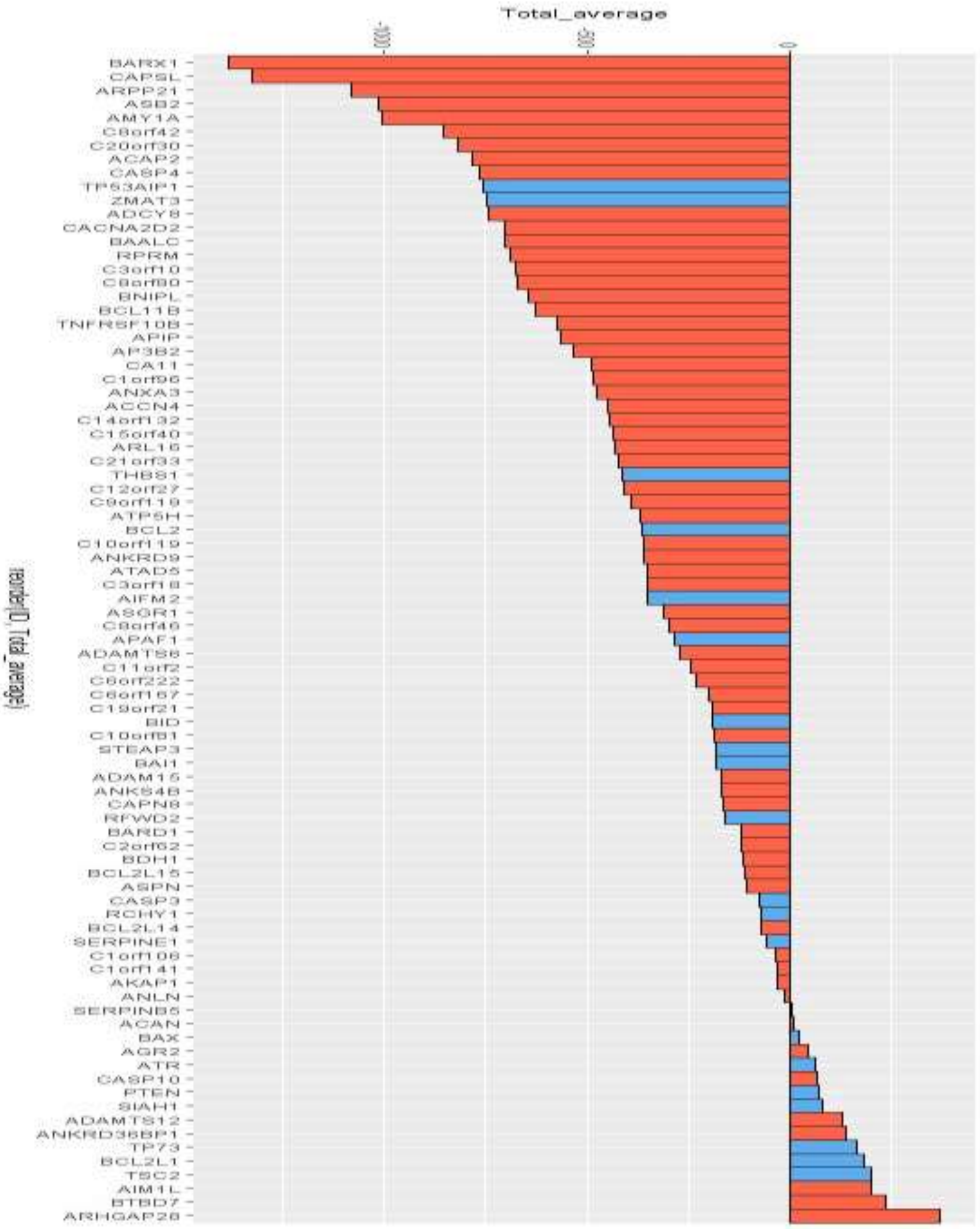


Figure 5.15: Differential drivers (as targets) associated with the TCGA-STAD-WTTP53 Sorted based on the total average of interactions. Blue colour indicates known TP53 pathway members, and red colour indicates novel targets

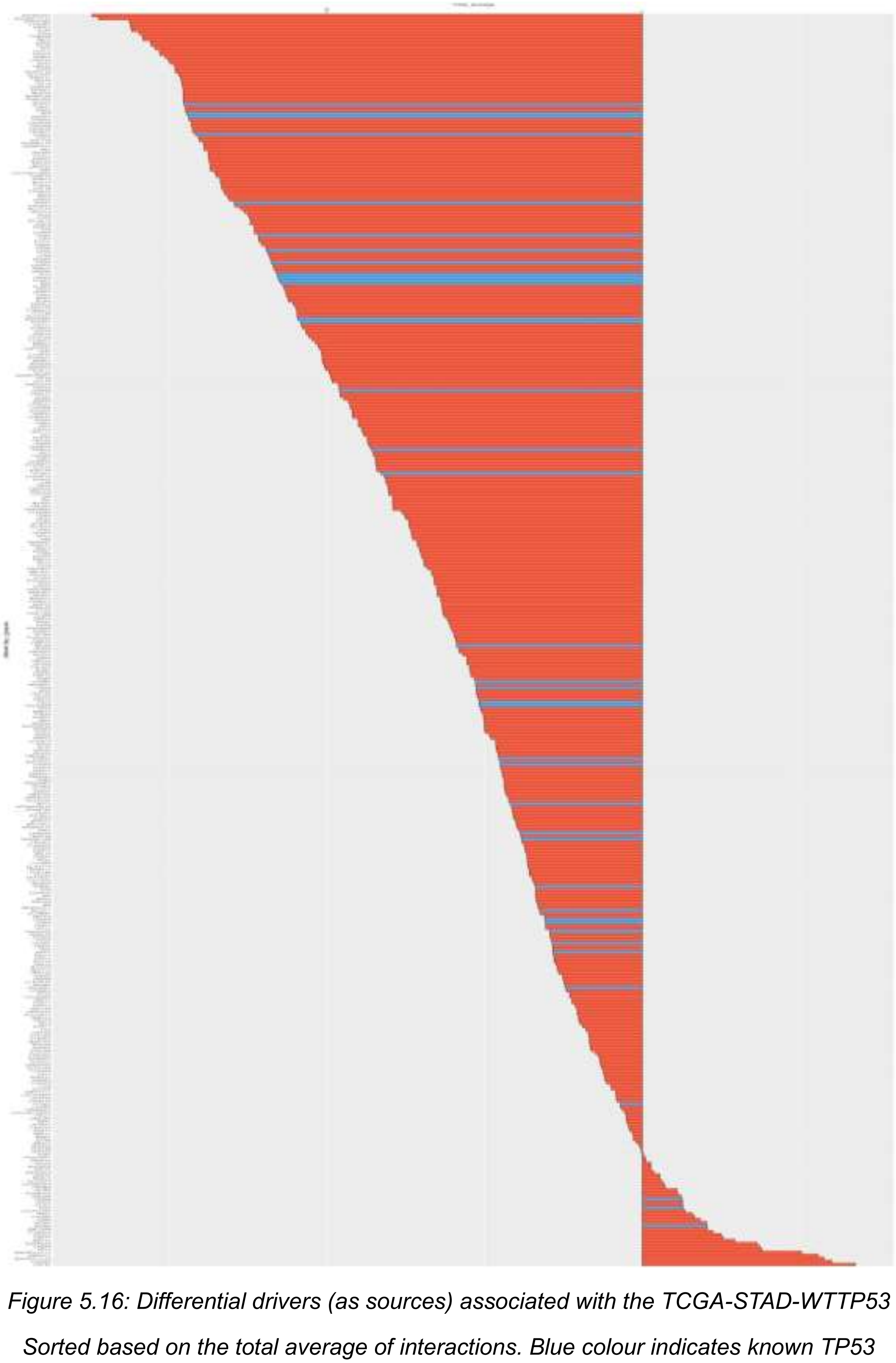


Figure 5.16: Differential drivers (as sources) associated with the TCGA-STAD-WTTP53 Sorted based on the total average of interactions. Blue colour indicates known TP53 pathway members, and red colour indicates novel targets

5.5.4. Combined analysis for the WTTP53 differential drivers of the three projects

The differential drivers associated with the TP53 pathway in the WTTP53 state for the investigated projects (COADREDA, PAAD, and STAD) were combined together to identify the concordant source drivers among all cohorts. The results indicate 11 novel source drivers, one of which (CENPM) was shared in the COADREAD and STAD projects, while the remaining 10 were common between the PAAD and STAD projects. Four of them (ITPR3, IQGAP3, EPS8L1 and GPRC5A) had high protein expression for the corresponding cancer type based on the Human Protein Atlas result. TP53 pathway members were also found to be concordant source drivers, including PIGS and CDK2. A parallel analysis was also done for the target drivers, and the results revealed five elements (EFNA4, CDCA2, AIM1L, ANXA3, and AP3B2). This outcome concurs with the source driver combined results (in terms of the five elements), which means that these genes were equality important as sources and as targets. Table 5.1 presents the result of the combined analysis of the source drivers.

Table 5.1: Combined analysis for the three sets (source drivers). Orange colour indicates novel sources, blue colour indicates known pathway members

Influencer (Sources)					
Gene Symbol	Gene Name	Project	Rank	Sum of Average	ABS
CENPM	Centromere Protein M	COADREAD	91	-73.6409	73.64087
		STAD	289	-92.6156	92.6156
MYEOV	Myeloma Overexpressed	PAAD	4	279.6271	279.6271
		STAD	395	-145.635	145.6347
ITPR3	Inositol 1,4,5-Trisphosphate Receptor Type 3	PAAD	15	230.6639	230.6639
		STAD	403	-148.243	148.2432
IQGAP3	IQ Motif Containing GTPase Activating Protein 3	PAAD	23	210.8181	210.8181
		STAD	386	-143.816	143.8164
EPS8L1	EPS8 Like 1	PAAD	28	202.7906	202.7906
		STAD	396	-145.689	145.6894
GPRC5A	G Protein-Coupled Receptor Class C Group 5 Member A	PAAD	36	190.4279	190.4279
		STAD	370	-137.063	137.0634
PLEK2	Pleckstrin 2	PAAD	59	158.324	158.324

Table 5.1: Combined analysis for the three sets (source drivers). Orange colour indicates novel sources, blue colour indicates known pathway members

Influencer (Sources)					
Gene Symbol	Gene Name	Project	Rank	Sum of Average	ABS
		STAD	285	-90.1002	90.10022
FOXQ1	Forkhead Box Q1	PAAD	62	156.4982	156.4982
		STAD	364	-133.574	133.5744
E2F8	E2F Transcription Factor 8	PAAD	63	154.5518	154.5518
		STAD	415	-159.575	159.5751
DEPDC1B	DEP Domain Containing 1B	PAAD	67	150.557	150.557
		STAD	402	-148.106	148.106
GRHL2	Grainyhead Like Transcription Factor 2	PAAD	80	134.9482	134.9482
		STAD	417	-162.312	162.3122
PIGS	Phosphatidylinositol Glycan Anchor Biosynthesis Class S	COADREAD	94	-78.3905	78.39048
		STAD	20	12.93958	12.93958
CDK2	Cyclin Dependent Kinase 2	PAAD	95	117.824	117.824
		STAD	327	-112.898	112.8976

5.6. Summary and conclusion

This chapter used ANNI approach to model the interactions between the distinctive predictors associated with the TP53 pathway in the WTTP53 state for each of the three TCGA projects (COADREDA, PAAD, and STAD). The results revealed key interactions and novel differential drivers for each of the investigated cohorts, involving seven major hub nodes associated with the pathway in the WTTP53 state for the COADREAD project, three of which (RPS27L, SNRPD1, and MPDU1) were novel to the TP53 pathway; the remaining four (SIAH1, ZMAT3, GADD45B, and Bax) were known pathway members.

RPS27L represents a novel subnetwork with the highest negative interactions. RPS27L protein was highly expressed in the WTTP53 based on the cBioPortal result. The interaction analysis for the TCGA-PAAD-WTTP53 cohort showed five main novel nodes (NXF2B, TMPRSS4, CDCA2, FM01, and RHBDL2). TMPRSS4 was also found to be one of the top target drivers, with the highest total number of interactions among the whole network. For the

TCGA-STAD-WTTP53 project interaction analysis, there were five novel hub targets associated with the pathway in the WTTP53 state: FUT8, SNORD116-26, TSNAXIP1, CHSY1, and TRAM1L1. FUT8 acted as a subnetwork, which was highly connected with other genes.

The combined analysis revealed the highest commonality between the differential source driver results of the PAAD and the STAD projects. Some of these features such as, MYEOV, FOXQ1, GRHL2 and ITPR3 have been linked previously to poor prognosis in pancreatic cancer based on EBI search (https://www.ebi.ac.uk/). However, most of these common elements are novel to the pathway. No previous research has relate them to the pathway. The presence of these drivers among the top hits in two out of three projects could suggest an important connection between these drivers and the pathway, It may also be an indication that the system is attempting to maintain balance by interacting with features that are not known to the pathway. In conclusion, this chapter reflected the power of the ANNI approach and driver analysis for quantifying and modelling the interactions within the TP53 pathway. It also identified novel drivers associated with the pathway in the Wild-type state of the TP53.

CHAPTER 6
GENERAL DISCUSSION AND CONCLUSION

6.1. Overview

Despite extensive progress in numerous research specialties over the decades, cancer remains a major global burden, with fast growth in incidence and mortality (Cui et al., 2020). Several genetic alterations are instrumental in cancer, involving diverse genes in related processes. Genetic abnormalities associated with cancer have been broadly documented since the early discovery of oncogenes and tumour suppressor genes, and signalling pathways are now recognized for their crucial roles in controlling cellular processes; they are therefore important in the development of cancer and its potential therapy (Yip & Papa, 2021). The *TP53* pathway plays a major role in controlling genomic stability and cell cycle progression. It contains a protein network of diverse inputs and downstream outputs which, upon activation (during cellular stress) lead to ultimate biological responses, such as cell cycle arrest and apoptosis. The pathway has an anti-proliferative role in response to various stresses. Thus it has been recognized as a tumour suppressor pathway. The *TP53* gene, which is the main regulator of the pathway, acts as a transcriptional factor that controls the biological function based on the stress signal.

The *TP53* gene is the most commonly mutated in human cancers, ranging from 30-50% prevalence in every cancer type, often at higher frequency in more advanced stages of cancer, with an overrepresentation of Missense mutations (Olivier et al., 2010). Increasing knowledge has been gained regarding the biology of *TP53* and its signalling mechanism, and the prognostic function of the Mutant *TP53* has been recognized as highly significant in cancer (Robles & Harris, 2010). Patients with *TP53* mutations usually have a worse prognosis, especially in certain common variants such as colorectal cancer and leukaemia. In breast cancer, it has been identified as an independent marker for poor prognosis (Petitjean at al., 2007). However, the clinical implications and development of effective therapeutic approaches remain challenging areas for emergent research. Although the information regarding the pathway elements and how they interact has been categorised and automated in the database, this knowledge is based on already published experimental results in the literature. It's possible that some details are lacking that are essential to understanding the pathway. For this, we looked at the entire *TP53* network by investigating both established and novel aspects of the *TP53* pathway in cancer. In order to find new drivers that could be added to the pathway,

In the field of targeted therapy, the mutation status of the *TP53* (either Mutant or Wild-type) affects the selection of therapeutic approaches. Degradation of the Mutant *TP53* and restoration/stabilization of the Wild-type *TP53* are among the approaches that have been used to restore the *TP53* pathway (Hernández Borrero & El-Deiry, 2021). The *TP53* mutation status with a focus on the Missense *TP53* mutation type was selected as the main subject for the analysis undertaken in this research.

6.2. Analytical approaches

A massive scientific effort has been made to understand the disease through the study of molecular alterations. High throughput technologies and muti-omics analysis, including microarray and sequencing approaches, have enabled a deeper insight into this complex disease by providing parallel analysis of multiple genes. This helps in improving the understanding and management of the disease. However, a complete and comprehensive understanding of cancer remains challenging since a huge amount of data has been generated as a result of omics technology application that requires new analytical approaches to make sense of omics data. Computational methods have contributed to the transformation of molecular genomic data into biologically relevant information that could be utilized in clinical settings to identify significant cancer patterns and use them to classify patients and provide therapeutic guidance. Particular approaches project new ways of data visualization based on pattern recognition, such as clustering, or class prediction, such as supervised classification. Moreover, a growing number of researchers are focused on network and pathway modelling to infer new genetic interactions and biological processes from expression data (Slonim, 2002).

A pathway level analysis can reduce the complexity and dimensionality of the gene expression data and thereby enhance analytical power by focusing on fewer tested hypotheses. Such analysis also facilities the interpretation of results based on the fact that genes that belong to a specific pathway are usually involved in certain characteristic biological functions (Zheng et al., 2020). There are multiple pathway analysis tools that have been discovered over the years to be used for the analysis of omics data. Some of these tools, such as pathway-level integrative approaches, use combined P-value to assess the statistical significance of each pathway among multiple sets. This could be predisposed to false estimation, especially for datasets with large sample sizes. Other techniques use existing knowledge from literature to construct a network for each pathway, which can be helpful in determining how the phenotypes differ significantly in their pathways, but does not permit novel discoveries.

Existent strategies utilizing machine learning approaches offer the advantage of incorporating multiple machine learning techniques, allowing them to classify samples according to the pathways that are strongly associated with the phenotype. However, they do not take into consideration the interaction between genes in the pathway, an issue that has been addressed in this research by developing a novel computational approach for pathway modelling based on ANN inference algorithm.

6.3. Study contributions

This study's ANN application was applied to analyse the *TP53* pathway in cancer, which enabled quantification and measurement of the interactions between members of the pathway. By evaluating existing features and suggesting novel ones that might be utilised to manage the pathway in cancer, the study deepens our understanding of the mechanism behind the development of cancer. ANN data mining technique, as explained previously, has the advantages of high statistical power, non-linearity, and low risk of false discoveries. Thus, it offers some novel advantages over commonly used existing methods. Also, the implementation of the ANNI algorithm adds strength to the method by efficiently model all the possible interactions between the identified genes. It also provides more information regarding the magnitude and the directionality of the interaction, and the application of the driver analysis enabled the identification of key molecular drivers and provided a ranking of these drivers based on their general influence upon the system of interest.

This study adds knowledge by referencing novel patterns and key drivers associated with the *TP53* pathway in three key areas:

By considering the analysis of the *TP53* pathway in one cancer type (colorectal cancer). By analysing the *TP53* pathway in three different cancers (colorectal, gastric, and pancreatic cancers) regarding certain conditions (the Mutant versus the Wild type status of the TP53 gene).

By comparing the obtained results of the second analysis against existing pathway analysis tool (MetaCore).

By validating the results using multiple datasets, each of which was analysed separately and independently, thereby reducing the errors that might occur and the risk of false discoveries.

Only consistent features across all cohorts were considered for the final results, and it is unlikely that such features would occur by random chance. Also, the selection of the data was undertaken for additional layers for validation of the results. The first analysis used transcriptomic datasets (E-MTAB-6698), which have been normalized and organized as a metadata set; the second analysis used high-quality RNA sequencing data from the TCGA.

Moreover, the consistency between the ANN and the MetaCore results represents an additional strength for the ANN analysis findings.

6.4. Study limitations

ANN data mining approaches have notable limitations, particularly regarding performance trade-offs and practicality. The computational timing required to generate results using the Stepwise ANN method is affected by the data size; greater data sizes (i.e., more voluminous data sets) take longer to analyse, entailing more costs. Based on the project’s broad timetable, this restricts the potential for future investigation. Another drawback is that ANN results are obtained in a large file format, which takes up a lot of storage space on the C drive, causing an additional computational burden. Due to the black box nature of the ANN algorithm, it is not completely clear how those results are obtained (as described in Chapter 2).

Additionally, since presenting all of the results is difficult, the emphasis was placed on those with high concordance (the top-200 outputs). Even though this might be a practical way to handle the results and take them a step further to interaction analysis, it limits the visualization of other outputs that could be potentially relevant. Furthermore, the complexity of the interaction matrices produces another limitation, making it crucial to select the right filtering strategy for the obtained results. As a result of selecting only the top-100 interactions for visualization, other relevant and consistent hubs may not be visible. The solution for this issue was the implementation of the driver analysis, whereby each input was assisted, for it is the general influence upon the others within a dataset.

Another issue is that experimental methods used to generate the data (microarray and RNA sequencing) also have limitations; these methods use different laboratory procedures, including probe selection, labelling, and hybridization procedures. In addition, researchers using different experimental conditions regarding sampling techniques, sample anatomical side, and patient ethnicity result in poor reproducibility between experiments. Although the data used for this study were carefully selected regarding the origin and the normalization technique used, it still entails some level of bias. Consequently, none of the modelled interactions are definitive or fully complete, and this must be considered when interpreting the outcomes of this study.

6.5. Recommendations for future research

Based on the limitations identified above, numerous future research directions can be discerned from the outcomes of this study. On the technical side, data analysis using ANN approaches for multiple omics data, including proteomics, epigenomics, and metabolomics, or using different cancer types and pathways could lead to a more comprehensive understanding of the data and may highlight its key patterns.

In this study, Stepwise ANN algorithm implementation was automated, which lead to a marked decrease in the processing time, with the possibility of considering more data. The automated path also included a section for results sorting, which also helped in reducing the required time and the technical errors that may occur due to manual sorting. However, full automation could be possible in future works. For instance, it may be possible to generate an automated pipeline to connect the Stepwise and the ANN inference algorithms. This would generate the results as a final output, reducing the storage capacity and computational burden required for analysis. Also, allows young scientists with little expertise to use AMM approaches and acquire new knowledge.

Another level of validation for the obtained results could be added by analysis of the data for a large cohort of patients, and further classification of the data for deeper technical insights is recommended. For instance, categorization of the phenotypes based on the clinical characteristics could lead to new knowledge that could be immediately useful in clinical settings. Moreover, experimental work also could add another direction for validation if the presented results get more scientific attention and find way toward clinical practice. This may include measuring the protein levels of selected hubs using laboratory techniques, such as IHC or q-PCR.

On the commercial side, future directions could include the presentation of the obtained ANN results in scientific conferences and company presentations to popularize ANN approaches (by demonstrating their utility and efficiency), thereby promoting possible external cooperation and increased buy-in from diverse stakeholders. For instance, the common differential source drivers associated with the TP53 pathway in the wild type mutation status presented in Chapter 5 could be used for potential future cooperation to develop common therapeutic strategy, requiring the presentation of the findings to the pharmaceutical community.

References

Abdel-Fatah, T. M., Powe, D. G., Agboola, J., Adamowicz-Brice, M., Blamey, R. W., LopezGarcia, M. A., Green, A. R., Reis-Filho, J. S., & Ellis, I. O. (2010). The biological, clinical and prognostic implications of p53 transcriptional pathways in breast cancers. *Journal of Pathology*, *220*(4), 419–434. https://doi.org/10.1002/path.2663

Abdel-Fatah, T. M., Ball, G., Chen, X., Mehaisi, D., Giannotti, E., Auer, D., ... & Chan, S. (2022, February). Utilising artificial intelligence (AI) for analysing multiplex genomic and magnetic resonance imaging (MRI) data to develop multimodality predictive system for personalised neoadjuvant treatment of breast cancer (BC). In CANCER RESEARCH (Vol. 82, No. 4). 615 CHESTNUT ST, 17TH FLOOR, PHILADELPHIA, PA 19106-4404 USA: AMER ASSOC CANCER RESEARCH.

Aittokallio, T., & Schwikowski, B. (2006). Graph-based methods for analysing networks in cell biology. *Briefings in Bioinformatics*, *7*(3), 243–255.

https://doi.org/10.1093/bib/bbl022

Almendro, V., Cheng, Y. K., Randles, A., Itzkovitz, S., Marusyk, A., Ametller, E., GonzalezFarre, X., Muñoz, M., Russnes, H. G., Helland, A., Rye, I. H., Borresen-Dale, A. L.,

Maruyama, R., van Oudenaarden, A., Dowsett, M., Jones, R. L., Reis-Filho, J.,

Gascon, P., Gönen, M., Michor, F., … Polyak, K. (2014). Inference of tumor evolution during chemotherapy by computational modeling and in situ analysis of genetic and phenotypic cellular diversity. *Cell Reports*, *6*(3), 514–527.

https://doi.org/10.1016/j.celrep.2013.12.041

Alves Martins, B. A., de Bulhões, G. F., Cavalcanti, I. N., Martins, M. M., de Oliveira, P. G., & Martins, A. M. A. (2019). Biomarkers in colorectal cancer: The role of translational proteomics research. *Frontiers in Oncology*, *9*, Article 1284.

https://doi.org/10.3389/fonc.2019.01284

Aubrey, B. J., Kelly, G. L., Janic, A., Herold, M. J., & Strasser, A. (2018). How does p53 induce apoptosis and how does this relate to p53-mediated tumour suppression?. *Cell Death and Differentiation*, *25*(1), 104–113.

https://doi.org/10.1038/cdd.2017.169

Babamohamadi, M., Babaei, E., Ahmed Salih, B., Babamohammadi, M., Jalal Azeez, H., & Othman, G. (2022). Recent findings on the role of wild-type and mutant p53 in cancer development and therapy. *Frontiers in Molecular Biosciences*, *9*, Article 903075. https://doi.org/10.3389/fmolb.2022.903075

Baker, S. J., Fearon, E. R., Nigro, J. M., Hamilton, S. R., Preisinger, A. C., Jessup, J. M., vanTuinen, P., Ledbetter, D. H., Barker, D. F., Nakamura, Y., White, R., & Vogelstein, B. (1989). Chromosome 17 deletions and p53 gene mutations in colorectal carcinomas. *Science*, *244*(4901), 217–221.

https://doi.org/10.1126/science.2649981

Bammler, T., Beyer, R. P., Bhattacharya, S., Boorman, G. A., Boyles, A., Bradford, B. U.,

Bumgarner, R. E., Bushel, P. R., Chaturvedi, K., Choi, D., Cunningham, M. L., Deng,

S., Dressman, H. K., Fannin, R. D., Farin, F. M., Freedman, J. H., Fry, R. C., Harper,

A., Humble, M. C., Hurban, P., … Members of the Toxicogenomics Research Consortium. (2005). Standardizing global gene expression analysis between laboratories and across platforms. *Nature Methods*, *2*(5), 351–356.

https://doi.org/10.1038/nmeth754

Bang, S., Jee, S., Son, H., Wi, Y. C., Kim, H., Park, H., Myung, J., Shin, S. J., & Paik, S. S. (2021). Loss of DUSP4 expression as a prognostic biomarker in clear cell renal cell carcinoma. *Diagnostics*, *11*(10), Article 1939.

https://doi.org/10.3390/diagnostics11101939

Bar, J., Moskovits, N., & Oren, M. (2010). Involvement of stromal p53 in tumor-stroma interactions. *Seminars in Cell & Developmental Biology*, *21*(1), 47–54.

https://doi.org/10.1016/j.semcdb.2009.11.006

Basheer, I. A., & Hajmeer, M. (2000). Artificial neural networks: Fundamentals, computing, design, and application. *Journal of Microbiological Methods*, *43*(1), 3–31.

https://doi.org/10.1016/s0167-7012(00)00201-3

Beggs, A. D., James, J., Caldwell, G., Prout, T., Dilworth, M. P., Taniere, P., Iqbal, T., Morton, D. G., & Matthews, G. (2018). Discovery and validation of methylation biomarkers for ulcerative colitis associated neoplasia. *Inflammatory Bowel Diseases*, *24*(7), 1503–1509. https://doi.org/10.1093/ibd/izy119

Bellman, R. (1961). *Adaptive control processes*. Princeton University Press.

Bhaskar, S., Singh, V. B., & Nayak, A. K. (2014). Managing data in SVM supervised algorithm for data mining technology. *2014 Conference on IT in Business, Industry and Government (CSIBIG)*, 1–4. http://dx.doi.org/10.1109/CSIBIG.2014.7056946 Bishop, C. M. (1995). *Neural networks for pattern recognition*. Oxford University Press.

Botía, J.A., Vandrovcova, J., Forabosco, P., Guelfi, S., D'Sa, K., United Kingdom Brain Expression Consortium, Hardy, J., Lewis, C.M., Ryten, M. and Weale, M.E. (2017). An additional k-means clustering step improves the biological features of WGCNA gene co-expression networks. *BMC systems biology*, 11, pp.1-16.

Brady, C. A., & Attardi, L. D. (2010). p53 at a glance. *Journal of Cell Science*, *123*(15), 2527–2532. https://doi.org/10.1242/jcs.064501

Brannon, A. R., Vakiani, E., Sylvester, B. E., Scott, S. N., McDermott, G., Shah, R. H.,

Kania, K., Viale, A., Oschwald, D. M., Vacic, V., Emde, A. K., Cercek, A., Yaeger, R.,

Kemeny, N. E., Saltz, L. B., Shia, J., D'Angelica, M. I., Weiser, M. R., Solit, D. B., & Berger, M. F. (2014). Comparative sequencing analysis reveals high genomic concordance between matched primary and metastatic colorectal cancer lesions. *Genome Biology*, *15*(8), Article 454. https://doi.org/10.1186/s13059-014-

0454-7

Brosh, R., & Rotter, V. (2009). When mutants gain new powers: News from the mutant p53 field. *Nature Reviews Cancer*, *9*(10), 701–713. https://doi.org/10.1038/nrc2693

Brown, M. P., Grundy, W. N., Lin, D., Cristianini, N., Sugnet, C. W., Furey, T. S., Ares, M. Jr., & Haussler, D. (2000). Knowledge-based analysis of microarray gene expression data by using support vector machines. *Proceedings of the National Academy of Sciences of the United States of America*, *97*(1), 262–267.

https://doi.org/10.1073/pnas.97.1.262

Chen, E. Y., Tan, C. M., Kou, Y., Duan, Q., Wang, Z., Meirelles, G. V., Clark, N. R., & Ma'ayan, A. (2013). Enrichr: Interactive and collaborative HTML5 gene list enrichment analysis tool. *BMC Bioinformatics*, *14*, Article 128.

https://doi.org/10.1186/1471-2105-14-128

Chen, X., Wang, M., & Zhang, H. (2011). The use of classification trees for bioinformatics. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, *1*(1), 55–63. https://doi.org/10.1002/widm.14

Chen, D., Zhong, N., Guo, Z., Ji, Q., Dong, Z., Zheng, J., ... & Song, T. (2023). MCM10, a potential diagnostic, immunological, and prognostic biomarker in pan-cancer. Scientific Reports, 13(1), 17701.

Chipman, H., & Tibshirani, R. (2006). Hybrid hierarchical clustering with applications to microarray data. *Biostatistics*, *7*(2), 286–301.

https://doi.org/10.1093/biostatistics/kxj007

Chung D. C. (2000). The genetic basis of colorectal cancer: Insights into critical pathways of tumorigenesis. *Gastroenterology*, *119*(3), 854–865.

https://doi.org/10.1053/gast.2000.16507

Collavin, L., Lunardi, A., & Del Sal, G. (2010). p53-family proteins and their regulators: Hubs and spokes in tumor suppression. *Cell Death and Differentiation*, *17*(6), 901–911. https://doi.org/10.1038/cdd.2010.35

Creixell, P., Reimand, J., Haider, S., Wu, G., Shibata, T., Vazquez, M., Mustonen, V., Gonzalez-Perez, A., Pearson, J., Sander, C., Raphael, B. J., Marks, D. S., Ouellette,

B. F. F., Valencia, A., Bader, G. D., Boutros, P. C., Stuart, J. M., Linding, R., Lopez-Bigas, N., Stein, L. D., … Mutation Consequences and Pathway Analysis Working Group of the International Cancer Genome Consortium. (2015). Pathway and network analysis of cancer genomes. *Nature Methods*, *12*(7), 615–621. https://doi.org/10.1038/nmeth.3440

Csikász-Nagy, A., Battogtokh, D., Chen, K. C., Novák, B., & Tyson, J. J. (2006). Analysis of a generic model of eukaryotic cell-cycle regulation. *Biophysical Journal*, *90*(12), 4361–4379. https://doi.org/10.1529/biophysj.106.081240

Cui, W., Aouidate, A., Wang, S., Yu, Q., Li, Y., & Yuan, S. (2020). Discovering anti-cancer drugs via computational methods. *Frontiers in Pharmacology*, *11*, Article 733. https://doi.org/10.3389/fphar.2020.00733

Demir, E., Cary, M. P., Paley, S., Fukuda, K., Lemer, C., Vastrik, I., Wu, G., D'Eustachio, P., Schaefer, C., Luciano, J., Schacherer, F., Martinez-Flores, I., Hu, Z., JimenezJacinto, V., Joshi-Tope, G., Kandasamy, K., Lopez-Fuentes, A. C., Mi, H., Pichler, E., Rodchenkov, I., … Bader, G. D. (2010). The BioPAX community standard for pathway data sharing. *Nature Biotechnology*, *28*(9), 935–942. https://doi.org/10.1038/nbt.1666

Dimitrakopoulos, C. M., & Beerenwinkel, N. (2017). Computational approaches for the identification of cancer genes and pathways. *Wiley Interdisciplinary Reviews. Systems Biology and Medicine*, *9*(1), Article e1364. https://doi.org/10.1002/wsbm.1364

Dong, W., Cui, J., Yang, J., Li, W., Wang, S., Wang, X., Li, X., Lu, Y., & Xiao, W. (2015). Decreased expression of Rab27A and Rab27B correlates with metastasis and poor prognosis in colorectal cancer. *Discovery Medicine*, *20*(112), 357–367.

Draghici, S., Khatri, P., Tarca, A. L., Amin, K., Done, A., Voichita, C., Georgescu, C., & Romero, R. (2007). A systems biology approach for pathway level analysis. *Genome Research*, *17*(10), 1537–1545. https://doi.org/10.1101/gr.6202607

Du, W., & Elemento, O. (2015). Cancer systems biology: Embracing complexity to develop better anticancer therapeutic strategies. *Oncogene*, *34*(25), 3215–3225. https://doi.org/10.1038/onc.2014.291

Efroni, S., Schaefer, C. F., & Buetow, K. H. (2007). Identification of key processes underlying cancer phenotypes using biologic pathway analysis. *PloS One*, *2*(5), Article e425. https://doi.org/10.1371/journal.pone.0000425

El-Deiry, W. S., Tokino, T., Velculescu, V. E., Levy, D. B., Parsons, R., Trent, J. M., Lin, D., Mercer, W. E., Kinzler, K. W., & Vogelstein, B. (1993). WAF1, a potential mediator of p53 tumor suppression. *Cell*, *75*(4), 817–825. https://doi.org/10.1016/00928674(93)90500-p

Fearon, E. R., & Vogelstein, B. (1990). A genetic model for colorectal tumorigenesis. *Cell*, *61*(5), 759–767. https://doi.org/10.1016/0092-8674(90)90186-i

Feng, Z., Hu, W., de Stanchina, E., Teresky, A. K., Jin, S., Lowe, S., & Levine, A. J. (2007). The regulation of AMPK beta1, TSC2, and PTEN expression by p53: Stress, cell and tissue specificity, and the role of these gene products in modulating the IGF-1-AKTmTOR pathways. *Cancer Research*, *67*(7), 3043–3053. https://doi.org/10.1158/00085472.CAN-06-4149

Finak, G., Bertos, N., Pepin, F., Sadekova, S., Souleimanova, M., Zhao, H., Chen, H., Omeroglu, G., Meterissian, S., Omeroglu, A., Hallett, M., & Park, M. (2008). Stromal gene expression predicts clinical outcome in breast cancer. *Nature Medicine*, *14*(5), 518–527. https://doi.org/10.1038/nm1764

Friedman, N., Linial, M., Nachman, I., & Pe'er, D. (2000). Using Bayesian networks to analyze expression data. *Journal of Computational Biology*, *7*(3–4), 601–620. https://doi.org/10.1089/106652700750050961

Gaiddon, C., Lokshin, M., Ahn, J., Zhang, T., & Prives, C. (2001). A subset of tumor-derived mutant forms of p53 down-regulate p63 and p73 through a direct interaction with the p53 core domain. *Molecular and Cellular Biology*, *21*(5), 1874–1887. https://doi.org/10.1128/MCB.21.5.1874-1887.2001

Gali-Muhtasib, H., Kuester, D., Mawrin, C., Bajbouj, K., Diestel, A., Ocker, M., Habold, C., Foltzer-Jourdainne, C., Schoenfeld, P., Peters, B., Diab-Assaf, M., Pommrich, U., Itani, W., Lippert, H., Roessner, A., & Schneider-Stock, R. (2008). Thymoquinone triggers inactivation of the stress response pathway sensor CHEK1 and contributes to apoptosis in colorectal cancer cells. *Cancer Research*, *68*(14), 5609–5618. https://doi.org/10.1158/0008-5472.CAN-08-0884

Gatenby R. (2012). Perspective: Finding cancer's first principles. *Nature*, *491*(7425), Article S55. https://doi.org/10.1038/491s55a.

Gao, Y., Zhao, H., Ren, M., Chen, Q., Li, J., Li, Z., ... & Yue, W. (2020). TOP2A promotes tumorigenesis of high-grade serous ovarian cancer by regulating the TGF-β/Smad pathway. Journal of Cancer, 11(14), 4181.

Grześ, M., & Kretowski, M. (2007). Decision tree approach to microarray data analysis. *Biocybernetics and Biomedical Engineering*, *27*(3), 29–42.

Gu, J., Huang, W., Zhang, J., Wang, X., Tao, T., Yang, L., Zheng, Y., Liu, S., Yang, J., Zhu, L., Wang, H., & Fan, Y. (2021). TMPRSS4 promotes cell proliferation and inhibits apoptosis in pancreatic ductal adenocarcinoma by activating ERK1/2 signaling pathway. *Frontiers in Oncology*, *11*, Article 628353. https://doi.org/10.3389/fonc.2021.628353

Guo, Q., Song, Y., Zhang, H., Wu, X., Xia, P., & Dang, C. (2013). Detection of hypermethylated fibrillin-1 in the stool samples of colorectal cancer patients. *Medical Oncology*, *30*(4), Article 695. https://doi.org/10.1007/s12032-013-0695-4

Guyon, I., & Elisseeff, A. (2003). An introduction to variable and feature selection. Journal of machine learning research, 3(Mar), 1157-1182.

Harper, J. W., Adami, G. R., Wei, N., Keyomarsi, K., & Elledge, S. J. (1993). The p21 Cdkinteracting protein Cip1 is a potent inhibitor of G1 cyclin-dependent kinases. *Cell*, *75*(4), 805–816. https://doi.org/10.1016/0092-8674(93)90499-g

Haykin, S. S. (2009). *Neural networks and learning machines* (3rd ed.). Pearson Education.

He, H., & Sun, Y. (2007). Ribosomal protein S27L is a direct p53 target that regulates apoptosis. *Oncogene*, *26*(19), 2707–2716. https://doi.org/10.1038/sj.onc.1210073

Hermeking, H., Lengauer, C., Polyak, K., He, T. C., Zhang, L., Thiagalingam, S., Kinzler, K. W., & Vogelstein, B. (1997). 14-3-3sigma is a p53-regulated inhibitor of G2/M progression. *Molecular Cell*, *1*(1), 3–11. https://doi.org/10.1016/s10972765(00)80002-7

Hernández Borrero, L. J., & El-Deiry, W. S. (2021). Tumor suppressor p53: Biology, signaling pathways, and therapeutic targeting. *Biochimica et Biophysica Acta. Reviews on Cancer*, *1876*(1), Article 188556. https://doi.org/10.1016/j.bbcan.2021.188556

Huang, C. J., Yang, S. H., Lee, C. L., Cheng, Y. C., Tai, S. Y., & Chien, C. C. (2013). Ribosomal protein S27-like in colorectal cancer: A candidate for predicting prognoses. *PloS One*, *8*(6), Article e67043. https://doi.org/10.1371/journal.pone.0067043

Huang, J. (2021). Current developments of targeting the p53 signaling pathway for cancer treatment. *Pharmacology & therapeutics*, 220, p.107720.

Irish, J. M., Hovland, R., Krutzik, P. O., Perez, O. D., Bruserud, Ø., Gjertsen, B. T., & Nolan, G. P. (2004). Single cell profiling of potentiated phospho-protein networks in cancer cells. *Cell*, *118*(2), 217–228. https://doi.org/10.1016/j.cell.2004.06.028

Jain, A.K., Mao, J., & Mohiuddin, K. M. (1996). Artificial neural networks: A tutorial. *Computer*, *29*(3), 31–44. https://doi.org/10.1109/2.485891

Kandoth, C., McLellan, M. D., Vandin, F., Ye, K., Niu, B., Lu, C., Xie, M., Zhang, Q., McMichael, J. F., Wyczalkowski, M. A., Leiserson, M. D. M., Miller, C. A., Welch, J. S., Walter, M. J., Wendl, M. C., Ley, T. J., Wilson, R. K., Raphael, B. J., & Ding, L. (2013). Mutational landscape and significance across 12 major cancer types. *Nature*, *502*(7471), 333–339. https://doi.org/10.1038/nature12634

Khatri, P., & Drăghici, S. (2005). Ontological analysis of gene expression data: Current tools, limitations, and open problems. *Bioinformatics*, *21*(18), 3587–3595. https://doi.org/10.1093/bioinformatics/bti565

Khatri, P., Sirota, M., & Butte, A. J. (2012). Ten years of pathway analysis: Current approaches and outstanding challenges. *PloS Computational Biology*, *8*(2), Article e1002375. https://doi.org/10.1371/journal.pcbi.1002375

Kim, E., Kim, J. Y., & Lee, J. Y. (2019). Mathematical modeling of p53 pathways. *International Journal of Molecular Sciences*, *20*(20), Article 5179.

https://doi.org/10.3390/ijms20205179

Kirouac, D. C., Du, J. Y., Lahdenranta, J., Overland, R., Yarar, D., Paragas, V., Pace, E.,

McDonagh, C. F., Nielsen, U. B., & Onsum, M. D. (2013). Computational modeling of ERBB2-amplified breast cancer identifies combined ErbB2/3 blockade as superior to the combination of MEK and AKT inhibitors. *Science Signaling*, *6*(288), Article ra68.

https://doi.org/10.1126/scisignal.2004008

Kohl, P., Noble, D., Winslow, R. L., & Hunter, P. J. (2000). Computational modelling of biological systems: Tools and visions. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, *358*(1766), 579–610.

https://doi.org/10.1098/rsta.2000.0547

Kuerbitz, S. J., Plunkett, B. S., Walsh, W. V., & Kastan, M. B. (1992). Wild-type p53 is a cell cycle checkpoint determinant following irradiation. *Proceedings of the National Academy of Sciences of the United States of America*, *89*(16), 7491–7495.

https://doi.org/10.1073/pnas.89.16.7491

Lancashire, L. J., Mian, S., Ellis, I. O., & Rees, R. C. (2005). Current developments in the analysis of proteomic data: Artificial neural network data mining techniques for the identification of proteomic biomarkers related to breast cancer. *Current Proteomics*, *200*(1), 15–29. http://dx.doi.org/10.2174/1570164053507808.

Lancashire, L.J., Rees, R.C. and Ball, G.R. (2008). Identification of gene transcript signatures predictive for estrogen receptor and lymph node status using a stepwise forward selection artificial neural network modelling approach. Artificial intelligence in medicine, 43(2), pp.99-111.

Lancashire, L.J., Lemetre, C. and Ball, G.R. (2009). An introduction to artificial neural networks in bioinformatics—application to complex microarray and mass spectrometry datasets in cancer studies. Briefings in bioinformatics, 10(3), pp.315329.

Lane, D. P., & Crawford, L. V. (1979). T antigen is bound to a host protein in SV40transformed cells. *Nature*, *278*(5701), 261–263. https://doi.org/10.1038/278261a0

Langerød, A., Zhao, H., Borgan, Ø., Nesland, J. M., Bukholm, I. R., Ikdahl, T., Kåresen, R., Børresen-Dale, A. L., & Jeffrey, S. S. (2007). TP53 mutation status and gene expression profiles are powerful prognostic markers of breast cancer. *Breast Cancer Research : BCR*, *9*(3), Article R30. https://doi.org/10.1186/bcr1675

Lemetre, C. (2010). *Artificial neural network techniques to investigate potential interactions between biomarkers* [Doctoral dissertation, Nottingham Trent University]. IRep. https://irep.ntu.ac.uk/id/eprint/142

Lemetre, C., Lancashire, L. J., Rees, R. C., & Ball, G. R. (2009). Artificial neural network based algorithm for biomolecular interactions modeling. In J. Cabestany, F. Sandoval, A. Prieto, & J. M. Corchado (Eds.), *Bio-inspired systems: Computational and ambient intelligence* (IWANN 2009, Lecture Notes in Computer Science, Vol. 5517). Springer. https://doi.org/10.1007/978-3-642-02478-8_110

Lev Bar-Or, R., Maya, R., Segel, L. A., Alon, U., Levine, A. J., & Oren, M. (2000). Generation of oscillations by the p53-Mdm2 feedback loop: A theoretical and experimental study. *Proceedings of the National Academy of Sciences of the United States of America*, *97*(21), 11250–11255. https://doi.org/10.1073/pnas.210171597

Liao, C., An, J., Yi, S., Tan, Z., Wang, H., Li, H., Guan, X., Liu, J., & Wang, Q. (2021). FUT8 and protein core fucosylation in tumours: From diagnosis to treatment. *Journal of Cancer*, *12*(13), 4109–4120. https://doi.org/10.7150/jca.58268 .

Libbrecht, M. W., & Noble, W. S. (2015). Machine learning applications in genetics and genomics. Nature Reviews Genetics, 16(6), 321-332.

Liu, B., Oltvai, Z. N., Bayır, H., Silverman, G. A., Pak, S. C., Perlmutter, D. H., & Bahar, I. (2017). Quantitative assessment of cell fate decision between autophagy and apoptosis. *Scientific Reports*, *7*(1), Article 17605. https://doi.org/10.1038/s41598-01718001-w

Liu, S., Yao, J., Li, H., Changpeng, Q., & Liu, R. (2019). Research on a method of fruit tree pruning based on BP neural network. *Journal of Physics Conference Series*, *1237*(4), Article 042047. http://dx.doi.org/10.1088/1742-6596/1237/4/042047

Louis, J. M., McFarland, V. W., May, P., & Mora, P. T. (1988). The phosphoprotein p53 is down-regulated post-transcriptionally during embryogenesis in vertebrates. *Biochimica et Biophysica Acta*, *950*(3), 395–402. https://doi.org/10.1016/0167-4781(88)90136-4

Lu, A. G., Feng, H., Wang, P. X., Han, D. P., Chen, X. H., & Zheng, M. H. (2012). Emerging roles of the ribonucleotide reductase M2 in colorectal cancer and ultraviolet-induced DNA damage repair. *World Journal of Gastroenterology*, *18*(34), 4704–4713. https://doi.org/10.3748/wjg.v18.i34.4704

Lu, X., Nguyen, T. A., Zhang, X., & Donehower, L. A. (2008). The Wip1 phosphatase and Mdm2: Cracking the “Wip” on p53 stability. *Cell Cycle*, *7*(2), 164–168. https://doi.org/10.4161/cc.7.2.5299

Madan Babu, M. M., Luscombe, N. M., Aravind, L., Gerstein, M., & Teichmann, S. A. (2004). Structure and evolution of transcriptional regulatory networks. *Current Opinion in Structural Biology*, *14*(3), 283–291. https://doi.org/10.1016/j.sbi.2004.05.004

Maden Babu, M. (2004). Introduction to microarray data analysis. In R. Grant (Ed.), *Computational genomics: Theory and application* (pp. 225–250). Taylor & Francis.

Manning, T., Sleator, R. D., & Walsh, P. (2014). Biologically inspired intelligent decision making: A commentary on the use of artificial neural networks in bioinformatics. *Bioengineered*, *5*(2), 80–95. https://doi.org/10.4161/bioe.26997

Mantovani, F., Collavin, L., & Del Sal, G. (2019). Mutant p53 as a guardian of the cancer cell. *Cell Death and Differentiation*, *26*(2), 199–212. https://doi.org/10.1038/s41418018-0246-9

May, P., & May, E. (1999). Twenty years of p53 research: Structural and functional aspects of the p53 protein. *Oncogene*, *18*(53), 7621–7636. https://doi.org/10.1038/sj.onc.1203285

Mayo, L. D., & Donner, D. B. (2001). A phosphatidylinositol 3-kinase/Akt pathway promotes translocation of Mdm2 from the cytoplasm to the nucleus. *Proceedings of the National Academy of Sciences of the United States of America*, *98*(20), 11598–11603. https://doi.org/10.1073/pnas.181181198

Mi, H., Muruganujan, A., Casagrande, J. T., & Thomas, P. D. (2013). Large-scale gene function analysis with the PANTHER classification system. *Nature Protocols*, *8*(8), 1551–1566. https://doi.org/10.1038/nprot.2013.092

Milner J. (1984). Different forms of p53 detected by monoclonal antibodies in non-dividing and dividing lymphocytes. *Nature*, *310*(5973), 143–145. https://doi.org/10.1038/310143a0

Mitchell, T. M. (1997). *Machine learning*. McGraw Hill.

Mitrea, C., Taghavi, Z., Bokanizad, B., Hanoudi, S., Tagett, R., Donato, M., Voichiţa, C., & Drăghici, S. (2013). Methods and approaches in the topology-based analysis of biological pathways. *Frontiers in Physiology*, *4*, Article 278. https://doi.org/10.3389/fphys.2013.00278

Muzio, G., O'Bray, L., & Borgwardt, K. (2021). Biological network analysis with deep learning. *Briefings in Bioinformatics*, *22*(2), 1515–1530. https://doi.org/10.1093/bib/bbaa257

Narrandes, S., & Xu, W. (2018). Gene expression detection assay for cancer clinical use. *Journal of Cancer*, *9*(13), 2249–2265. https://doi.org/10.7150/jca.24744

Nguyen, H.T. and Duong, H.Q. (2018). The molecular characteristics of colorectal cancer: Implications for diagnosis and therapy. Oncology letters, 16(1), pp.9-18.

Nguyen, T. M., Shafi, A., Nguyen, T., & Draghici, S. (2019). Identifying significantly impacted pathways: A comprehensive review and assessment. *Genome Biology*, *20*(1), Article 203. https://doi.org/10.1186/s13059-019-1790-4

Ohashi, T., Idogawa, M., Sasaki, Y., & Tokino, T. (2017). p53 mediates the suppression of cancer cell invasion by inducing LIMA1/EPLIN. *Cancer Letters*, *390*, 58–66. https://doi.org/10.1016/j.canlet.2016.12.034

Oliner, J. D., Kinzler, K. W., Meltzer, P. S., George, D. L., & Vogelstein, B. (1992). Amplification of a gene encoding a p53-associated protein in human sarcomas. *Nature*, *358*(6381), 80–83. https://doi.org/10.1038/358080a0

Olivier, M., Hollstein, M., & Hainaut, P. (2010). TP53 mutations in human cancers: Origins, consequences, and clinical use. *Cold Spring Harbor Perspectives in Biology*, *2*(1), Article a001008. https://doi.org/10.1101/cshperspect.a001008

Papin, J. A., Hunter, T., Palsson, B. O., & Subramaniam, S. (2005). Reconstruction of cellular signalling networks and analysis of their properties. *Nature Reviews Molecular Cell Biology*, *6*(2), 99–111. https://doi.org/10.1038/nrm1570

Paterson, C., Clevers, H., & Bozic, I. (2020). Mathematical model of colorectal cancer initiation. *Proceedings of the National Academy of Sciences of the United States of America*, *117*(34), 20681–20688. https://doi.org/10.1073/pnas.2003771117

Pe'er, D., & Hacohen, N. (2011). Principles and strategies for developing network models in cancer. *Cell*, *144*(6), 864–873. https://doi.org/10.1016/j.cell.2011.03.001

Petitjean, A., Achatz, M. I., Borresen-Dale, A. L., Hainaut, P., & Olivier, M. (2007). TP53 mutations in human cancers: Functional selection and impact on cancer prognosis and outcomes. *Oncogene*, *26*(15), 2157–2165. https://doi.org/10.1038/sj.onc.1210302

Pirim, H., Ekşioğlu, B., Perkins, A., & Yüceer, C. (2012). Clustering of high throughput gene expression data. *Computers & Operations Research*, *39*(12), 3046–3061. https://doi.org/10.1016/j.cor.2012.03.008

Purvis, J. E., Karhohs, K. W., Mock, C., Batchelor, E., Loewer, A., & Lahav, G. (2012). p53 dynamics control cell fate. *Science*, *336*(6087), 1440–1444. https://doi.org/10.1126/science.1218351

Quackenbush J. (2001). Computational analysis of microarray data. *Nature Reviews Genetics*, *2*(6), 418–427. https://doi.org/10.1038/35076576

Rambau, P. F., Odida, M., & Wabinga, H. (2008). p53 expression in colorectal carcinoma in relation to histopathological features in Ugandan patients. *African Health Sciences*, *8*(4), 234–238.

Ratsada, P., Hijiya, N., Hidano, S., Tsukamoto, Y., Nakada, C., Uchida, T., ... & Moriyama, M. (2020). DUSP4 is involved in the enhanced proliferation and survival of DUSP4-

overexpressing cancer cells. Biochemical and Biophysical Research Communications, 528(3), 586-593.

Rauber, A., Merkl, D., & Dittenbach, M. (2002). The growing hierarchical self-organizing map: exploratory analysis of high-dimensional data. IEEE Transactions on Neural Networks, 13(6), 1331-1341.

Rawla, P., Sunkara, T., & Barsouk, A. (2019). Epidemiology of colorectal cancer: Incidence, mortality, survival, and risk factors. *Przeglad Gastroenterologiczny*, *14*(2), 89–103. https://doi.org/10.5114/pg.2018.81072

Reich, N. C., Oren, M., & Levine, A. J. (1983). Two distinct mechanisms regulate the levels of a cellular tumor antigen, p53. *Molecular and Cellular Biology*, *3*(12), 2143–2150. https://doi.org/10.1128/mcb.3.12.2143-2150.1983

Ringnér, M., Peterson, C., & Khan, J. (2002). Analyzing array data using supervised methods. *Pharmacogenomics*, *3*(3), 403–415. https://doi.org/10.1517/14622416.3.3.403

Ren, Z.J., Zhao, Y., Wang, G., Zhang, Z.C., Ma, L., Teng, M.Z. and Li, Y.M. (2022). Identification of differentially expressed miRNAs derived from serum exosomes associated with gastric cancer by microarray analysis. *Clinica Chimica Acta*, 531, pp.25-35.

Rezende, P.M., Xavier, J.S., Ascher, D.B., Fernandes, G.R. and Pires, D.E. (2022). Evaluating hierarchical machine learning approaches to classify biological databases. *Briefings in Bioinformatics,* 23(4).

Ro, S. H., Xue, X., Ramakrishnan, S. K., Cho, C. S., Namkoong, S., Jang, I., Semple, I. A., Ho, A., Park, H. W., Shah, Y. M., & Lee, J. H. (2016). Tumor suppressive role of sestrin2 during colitis and colon carcinogenesis. *Elife*, *5*, Article e12204. https://doi.org/10.7554/eLife.12204

Robles, A. I., & Harris, C. C. (2010). Clinical outcomes and correlates of TP53 mutations and cancer. *Cold Spring Harbor Perspectives in Biology*, *2*(3), Article a001016. https://doi.org/10.1101/cshperspect.a001016

Russo, G., Zegar, C., & Giordano, A. (2003). Advantages and limitations of microarray technology in human cancer. *Oncogene*, *22*(42), 6497–6507. https://doi.org/10.1038/sj.onc.1206865

Salarpour, F., Goudarzipour, K., Mohammadi, M.H., Ahmadzadeh, A., Faraahi, S., Allahbakhshian, A. and Farsani, M.A. (2020). Evaluation of growth factor independence 1 expression in patients with de novo acute myeloid leukemia. *Journal of Cancer Research and Therapeutics,* 16(1), pp.23-27.

Segal, E., Shapira, M., Regev, A., Pe'er, D., Botstein, D., Koller, D., & Friedman, N. (2003). Module networks: Identifying regulatory modules and their condition-specific regulators from gene expression data. *Nature Genetics*, *34*(2), 166–176.

https://doi.org/10.1038/ng1165

Shao, J. (1993). Linear model selection by cross-validation. *Journal of the American Statistical Association*, *88*(442), 486–494.

https://doi.org/10.1080/01621459.1993.10476299

Shi, D., & Gu, W. (2012). Dual roles of MDM2 in the regulation of p53: Ubiquitination dependent and ubiquitination independent mechanisms of MDM2 repression of p53 activity. *Genes & Cancer*, *3*(3–4), 240–248.

https://doi.org/10.1177/1947601912455199

Simon, R., Radmacher, M. D., Dobbin, K., & McShane, L. M. (2003). Pitfalls in the use of DNA microarray data for diagnostic and prognostic classification. *Journal of the National Cancer Institute*, *95*(1), 14–18. https://doi.org/10.1093/jnci/95.1.14

Slattery, M. L., Mullany, L. E., Wolff, R. K., Sakoda, L. C., Samowitz, W. S., & Herrick, J. S. (2019). The p53-signaling pathway and colorectal cancer: Interactions between downstream p53 target genes and miRNAs. *Genomics*, *111*(4), 762–771.

https://doi.org/10.1016/j.ygeno.2018.05.006

Slonim D. K. (2002). From patterns to pathways: Gene expression data analysis comes of age. *Nature Genetics*, *32*, 502–508. https://doi.org/10.1038/ng1033

Song, X., Zhou, L., Yang, W., Li, X., Ma, J., Qi, K., ... & Liang, B. (2023). PHLDA1 is a P53 target gene involved in P53-mediated cell apoptosis. Molecular and Cellular Biochemistry, 1-12.

Subramanian, A., Tamayo, P., Mootha, V. K., Mukherjee, S., Ebert, B. L., Gillette, M. A., Paulovich, A., Pomeroy, S. L., Golub, T. R., Lander, E. S., & Mesirov, J. P. (2005). Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles. *Proceedings of the National Academy of Sciences of the United States of America*, *102*(43), 15545–15550.

https://doi.org/10.1073/pnas.0506580102

Sun, T., Yang, W., Liu, J., & Shen, P. (2011). Modeling the basal dynamics of p53 system. *PloS One*, *6*(11), Article e27882.

https://doi.org/10.1371/journal.pone.0027882

Takekawa, M., Adachi, M., Nakahata, A., Nakayama, I., Itoh, F., Tsukuda, H., Taya, Y., & Imai, K. (2000). p53-inducible wip1 phosphatase mediates a negative feedback regulation of p38 MAPK-p53 signaling in response to UV radiation. *The EMBO Journal*, *19*(23), 6517–6526. https://doi.org/10.1093/emboj/19.23.6517

Tan, Y. S., Mhoumadi, Y., & Verma, C. S. (2019). Roles of computational modelling in understanding p53 structure, biology, and its therapeutic targeting. *Journal of Molecular Cell Biology*, *11*(4), 306–316. https://doi.org/10.1093/jmcb/mjz009

Teodoro, J. G., Evans, S. K., & Green, M. R. (2007). Inhibition of tumor angiogenesis by p53: A new role for the guardian of the genome. *Journal of Molecular Medicine*, *85*(11), 1175–1186. https://doi.org/10.1007/s00109-007-0221-2

Tian, X., Huang, B., Zhang, X. P., Lu, M., Liu, F., Onuchic, J. N., & Wang, W. (2017). Modeling the response of a tumor-suppressive network to mitogenic and oncogenic signals. *Proceedings of the National Academy of Sciences of the United States of America*, *114*(21), 5337–5342. https://doi.org/10.1073/pnas.1702412114

Tinker, A. V., Boussioutas, A., & Bowtell, D. D. (2006). The challenges of gene expression microarrays for the study of human cancer. *Cancer Cell*, *9*(5), 333–339. https://doi.org/10.1016/j.ccr.2006.05.001

Todorov, H., Fournier, D., & Gerber, S. (2018). Principal components analysis: Theory and application to gene expression data analysis. *Genomics and Computational Biology*, *4*(2), Article 100041. http://dx.doi.org/10.18547/gcb.2018.vol4.iss2.e100041

Toettcher, J. E., Loewer, A., Ostheimer, G. J., Yaffe, M. B., Tidor, B., & Lahav, G. (2009). Distinct mechanisms act in concert to mediate cell cycle arrest. *Proceedings of the National Academy of Sciences of the United States of America*, *106*(3), 785–790. https://doi.org/10.1073/pnas.0806196106

Torre, L. A., Siegel, R. L., Ward, E. M., & Jemal, A. (2016). Global cancer incidence and mortality rates and trends: An update. *Cancer Epidemiology, Biomarkers & Prevention*, *25*(1), 16–27. https://doi.org/10.1158/1055-9965.EPI-15-0578

Tuncbag, N., Kar, G., Gursoy, A., Keskin, O., & Nussinov, R. (2009). Towards inferring time dimensionality in protein-protein interaction networks by integrating structures: The p53 example. *Molecular Biosystems*, *5*(12), 1770–1778. https://doi.org/10.1039/B905661K

Urist, M., & Prives, C. (2002). p53 leans on its siblings. *Cancer Cell*, *1*(4), 311–313. https://doi.org/10.1016/s1535-6108(02)00064-8

Vacante, M., Borzì, A. M., Basile, F., & Biondi, A. (2018). Biomarkers in colorectal cancer: Current clinical utility and future perspectives. *World Journal of Clinical Cases*, *6*(15), 869–881. https://doi.org/10.12998/wjcc.v6.i15.869

Vaske, C. J., Benz, S. C., Sanborn, J. Z., Earl, D., Szeto, C., Zhu, J., Haussler, D., & Stuart, J. M. (2010). Inference of patient-specific pathway activities from multi-dimensional cancer genomics data using PARADIGM. *Bioinformatics*, *26*(12), i237–i245. https://doi.org/10.1093/bioinformatics/btq182

Villaamil, V. M., Gallego, G. A., Caínzos, I. S., Ruvira, L. V., Valladares-Ayerbes, M., & Aparicio, L. M. (2011). Relevant networks involving the p53 signalling pathway in renal cell carcinoma. *International Journal of Biomedical Science*, *7*(4), 273–282.

Vousden, K. H., & Lu, X. (2002). Live or let die: The cell's response to p53. *Nature Reviews. Cancer*, *2*(8), 594–604. https://doi.org/10.1038/nrc864

Vousden, K. H., & Prives, C. (2009). Blinded by the light: The growing complexity of p53. *Cell*, *137*(3), 413–431. https://doi.org/10.1016/j.cell.2009.04.037

Vousden, K. H., & Ryan, K. M. (2009). p53 and metabolism. *Nature Reviews Cancer*, *9*(10), 691–700. https://doi.org/10.1038/nrc2715.

Wang, Z., Gerstein, M., & Snyder, M. (2009). RNA-Seq: a revolutionary tool for transcriptomics. *Nature reviews genetics*, 10(1), 57-63.

Wang, Q. S., Shi, L. L., Sun, F., Zhang, Y. F., Chen, R. W., Yang, S. L., & Hu, J. L. (2019). High expression of *Anxa2* pseudogene *Anxa2p2* promotes an aggressive phenotype in hepatocellular carcinoma. *Disease Markers*, Article 9267046. https://doi.org/10.1155/2019/9267046

Wuest, T., Weimer, D., Irgens, C. and Thoben, K.D. (2016). Machine learning in manufacturing: advantages, challenges, and applications. *Production & Manufacturing Research*, 4(1), pp.23-45.

Xu, R., & Wunsch, D. C. (2010). Clustering algorithms in biomedical research: A review. *IEEE Reviews in Biomedical Engineering*, *3*, 120–154. https://doi.org/10.1109/RBME.2010.2083647

Yang, C., Li, D., Bai, Y., Song, S., Yan, P., Wu, R., Zhang, Y., Hu, G., Lin, C., Li, X., & Huang, L. (2018). DEAD-box helicase 27 plays a tumor-promoter role by regulating the stem cell-like activity of human colorectal cancer cells. *Oncotargets and Therapy*, *12*, 233–241. https://doi.org/10.2147/OTT.S190814

Yang, H. Y., Wen, Y. Y., Chen, C. H., Lozano, G., & Lee, M. H. (2003). 14-3-3 sigma positively regulates p53 and suppresses tumor growth. *Molecular and Cellular Biology*, *23*(20), 7096–7107. https://doi.org/10.1128/MCB.23.20.7096-7107.2003

Yeung, K. Y., & Ruzzo, W. L. (2001). Principal component analysis for clustering gene expression data. *Bioinformatics*, *17*(9), 763–774. https://doi.org/10.1093/bioinformatics/17.9.763

Yip, H. Y. K., & Papa, A. (2021). Signaling pathways in cancer: Therapeutic targets, combinatorial treatments, and new developments. *Cells*, *10*(3), Article 659. https://doi.org/10.3390/cells10030659

Yong, L., YuFeng, Z., & Guang, B. (2018). Association between PPP2CA expression and colorectal cancer prognosis tumor marker prognostic study. *International Journal of Surgery*, *59*, 80–89. https://doi.org/10.1016/j.ijsu.2018.09.020

Zafeiris, D., Rutella, S., & Ball, G. R. (2018). An artificial neural network integrated pipeline for biomarker discovery using Alzheimer's disease as a case study. Computational and Structural Biotechnology Journal, 16, 77-87

Zhang, B., Tian, Y., & Zhang, Z. (2014). Network biology in medicine and beyond. *Circulation. Cardiovascular Genetics*, *7*(4), 536–547.

https://doi.org/10.1161/CIRCGENETICS.113.000123

Index

Appendix

Complete datasets and result tables used in the project have been moved to the University OneDrive cloud store. These files will not be added to this manuscript due to their large size and can be accessed if required through contacting the project supervisor.

Complete TP53 pathway gene list from KEGG database

initial_alias	description
TP53	tumor protein p53 [Source:HGNC Symbol;Acc:HGNC:11998]
PTEN	phosphatase and tensin homolog [Source:HGNC Symbol;Acc:HGNC:9588]
CD82	CD82 molecule [Source:HGNC Symbol;Acc:HGNC:6210]
SESN1	sestrin 1 [Source:HGNC Symbol;Acc:HGNC:21595]
TSC2	TSC complex subunit 2 [Source:HGNC Symbol;Acc:HGNC:12363]
THBS1	thrombospondin 1 [Source:HGNC Symbol;Acc:HGNC:11785]
SERPINE1	serpin family E member 1 [Source:HGNC Symbol;Acc:HGNC:8583]
ADGRB1	adhesion G protein-coupled receptor B1 [Source:HGNC Symbol;Acc:HGNC:943]
SERPINB5	serpin family B member 5 [Source:HGNC Symbol;Acc:HGNC:8949]
DDB2	damage specific DNA binding protein 2 [Source:HGNC Symbol;Acc:HGNC:2718]
RRM2B	ribonucleotide reductase regulatory TP53 inducible subunit M2B [Source:HGNC Symbol;Acc:HGNC:17296]
RCHY1	ring finger and CHY zinc finger domain containing 1 [Source:HGNC Symbol;Acc:HGNC:17479]
CDK4	cyclin dependent kinase 4 [Source:HGNC Symbol;Acc:HGNC:1773]
CCNG1	cyclin G1 [Source:HGNC Symbol;Acc:HGNC:1592]
STEAP3	STEAP3 metalloreductase [Source:HGNC Symbol;Acc:HGNC:24592]
RFWD2	None
TP73	tumor protein p73 [Source:HGNC Symbol;Acc:HGNC:12003]
PPM1D	protein phosphatase, Mg2+/Mn2+ dependent 1D [Source:HGNC Symbol;Acc:HGNC:9277]
APAF1	apoptotic peptidase activating factor 1 [Source:HGNC Symbol;Acc:HGNC:576]
BAX	BCL2 associated X, apoptosis regulator [Source:HGNC Symbol;Acc:HGNC:959]
CASP3	caspase 3 [Source:HGNC Symbol;Acc:HGNC:1504]
CASP9	caspase 9 [Source:HGNC Symbol;Acc:HGNC:1511]
CDK2	cyclin dependent kinase 2 [Source:HGNC Symbol;Acc:HGNC:1771]
FAS	Fas cell surface death receptor [Source:HGNC Symbol;Acc:HGNC:11920]
IGFBP3	insulin like growth factor binding protein 3 [Source:HGNC Symbol;Acc:HGNC:5472]
CDKN2A	cyclin dependent kinase inhibitor 2A [Source:HGNC Symbol;Acc:HGNC:1787]

CCNE1	cyclin E1 [Source:HGNC Symbol;Acc:HGNC:1589]
TNFRSF10B	TNF receptor superfamily member 10b [Source:HGNC Symbol;Acc:HGNC:11905]
GADD45G	growth arrest and DNA damage inducible gamma [Source:HGNC Symbol;Acc:HGNC:4097]
GTSE1	G2 and S-phase expressed 1 [Source:HGNC Symbol;Acc:HGNC:13698]
PERP	p53 apoptosis effector related to PMP22 [Source:HGNC Symbol;Acc:HGNC:17637]
PIDD1	p53-induced death domain protein 1 [Source:HGNC Symbol;Acc:HGNC:16491]
PIGS	phosphatidylinositol glycan anchor biosynthesis class S [Source:HGNC Symbol;Acc:HGNC:14937]
CDK1	cyclin dependent kinase 1 [Source:HGNC Symbol;Acc:HGNC:1722]
CCND1	cyclin D1 [Source:HGNC Symbol;Acc:HGNC:1582]
CDK6	cyclin dependent kinase 6 [Source:HGNC Symbol;Acc:HGNC:1777]
CDKN1A	cyclin dependent kinase inhibitor 1A [Source:HGNC Symbol;Acc:HGNC:1784]
CHEK2	checkpoint kinase 2 [Source:HGNC Symbol;Acc:HGNC:16627]
AIFM2	apoptosis inducing factor mitochondria associated 2 [Source:HGNC Symbol;Acc:HGNC:21411]
ATM	ATM serine/threonine kinase [Source:HGNC Symbol;Acc:HGNC:795]
BCL2	BCL2 apoptosis regulator [Source:HGNC Symbol;Acc:HGNC:990]
BCL2L1	BCL2 like 1 [Source:HGNC Symbol;Acc:HGNC:992]
BID	BH3 interacting domain death agonist [Source:HGNC Symbol;Acc:HGNC:1050]
CASP8	caspase 8 [Source:HGNC Symbol;Acc:HGNC:1509]
CCNB1	cyclin B1 [Source:HGNC Symbol;Acc:HGNC:1579]
CYCS	cytochrome c, somatic [Source:HGNC Symbol;Acc:HGNC:19986]
GADD45B	growth arrest and DNA damage inducible beta [Source:HGNC Symbol;Acc:HGNC:4096]
GADD45A	growth arrest and DNA damage inducible alpha [Source:HGNC Symbol;Acc:HGNC:4095]
IGF1	insulin like growth factor 1 [Source:HGNC Symbol;Acc:HGNC:5464]
PMAIP1	phorbol-12-myristate-13-acetate-induced protein 1 [Source:HGNC Symbol;Acc:HGNC:9108]
ROS1	ROS proto-oncogene 1, receptor tyrosine kinase [Source:HGNC Symbol;Acc:HGNC:10261]
SHISA5	shisa family member 5 [Source:HGNC Symbol;Acc:HGNC:30376]
MDM4	MDM4 regulator of p53 [Source:HGNC Symbol;Acc:HGNC:6974]
MDM2	MDM2 proto-oncogene [Source:HGNC Symbol;Acc:HGNC:6973]
RPRM	reprimo, TP53 dependent G2 arrest mediator homolog [Source:HGNC Symbol;Acc:HGNC:24201]

SFN	stratifin [Source:HGNC Symbol;Acc:HGNC:10773]
SIAH1	siah E3 ubiquitin protein ligase 1 [Source:HGNC Symbol;Acc:HGNC:10857]
CHEK1	checkpoint kinase 1 [Source:HGNC Symbol;Acc:HGNC:1925]
SIVA1	SIVA1 apoptosis inducing factor [Source:HGNC Symbol;Acc:HGNC:17712]
TP53AIP1	tumor protein p53 regulated apoptosis inducing protein 1 [Source:HGNC Symbol;Acc:HGNC:29984]
ZMAT3	zinc finger matrin-type 3 [Source:HGNC Symbol;Acc:HGNC:29983]
ATR	ATR serine/threonine kinase [Source:HGNC Symbol;Acc:HGNC:882]

Commonality Tables for Data Used in Chapter 4 (A)

GSE13294		**GSE17536**		**GSE26682**	
Gene Name	Commonalities for all pathway members	Gene Name	Commonalities for all pathway members	Gene Name	Commonalities for all pathway members
DDX27	11	MND1	14	MRPL42	18
ATP9A	10	BUB1	12	RAN	18
CBFA2T2	10	CEP55	12	IER3IP1	18
RNF19B	10	MELK	12	MDH1	17
ABHD3	9	CDCA5	11	NDUFAB1	17
CDC45	9	KIF11	11	ATP5B	17
HSPA4L	9	MAD2L1	11	HAT1	17
MCM10	9	TNFSF9	11	ELAVL1	16
MED31	9	BIRC5	10	MRPL44	16
RAB27B	9	CDC20	10	MRPS18C	16
RPL22L1	9	CDC45	10	PSMD8	16
ALYREF	8	CDK1	10	C1QBP	15
BIRC5	8	PLK2	10	C2orf47	15
C18orf25	8	RAD51AP1	10	GSPT1	15
CACYBP	8	RRM1	10	H2AFZ	15
CCDC68	8	SHCBP1	10	MCMBP	15
CDCA5	8	TRIP13	10	MRPL11	15
CENPE	8	ZWINT	10	PSMB5	15
DNAJC9	8	ANLN	9	STOML2	15

GSE13294		**GSE17536**		**GSE26682**	
Gene Name	Commonalities for all pathway members	Gene Name	Commonalities for all pathway members	Gene Name	Commonalities for all pathway members
FDXR	8	AURKB	9	TRAPPC8	15
FEN1	8	BUB1B	9	JKAMP	15
G3BP2	8	CASC5	9	RBM12	15
HPSE	8	CCNB1	9	NDUFB3	15
MAP2K1	8	CCNB2	9	SNRPD1	15
MND1	8	DNAJC9	9	COX8A	14
PGAM1	8	DTL	9	CPSF6	14
PGM2	8	ECT2	9	FBXO22	14
RAN	8	FAS	9	GHITM	14
RFC4	8	FEN1	9	HNRNPF	14
SCO2	8	H2AFZ	9	MRPS11	14
TIFA	8	KIF14	9	MRPS12	14
TIMELESS	8	KIF23	9	MRPS16	14
TYMP	8	MCM10	9	NUP37	14
ANKFY1	7	OIP5	9	PGAM1	14
APOL2	7	ORC1	9	PSMA1	14
AURKB	7	ORC6	9	UQCRFS1	14
BLOC1S2	7	PBK	9	PHLDB1	14
BUB1B	7	PLK1	9	PRB1	14
C12orf57	7	PRC1	9	NDUFA12	14
C18orf8	7	RACGAP1	9	PDIA6	14
CAND1	7	RAD51	9	PPP1CC	14
CCNB1	7	SHMT2	9	CRK	14
DDIAS	7	SOCS6	9	MAPK1	14

GSE13294		GSE17536		GSE26682	
Gene Name	Commonalities for all pathway members	Gene Name	Commonalities for all pathway members	Gene Name	Commonalities for all pathway members
DDX46	7	TIMELESS	9	PTGES3	14
E2F7	7	TK1	9	SARNP	14
FANCI	7	UBE2S	9	ACP1	13
FBXO5	7	ARL4C	8	ADK	13
FBXO8	7	BLOC1S2	8	CCNB1	13
FECH	7	CCNA2	8	DCUN1D5	13
FOXM1	7	CDC6	8	EEF1E1	13
FTX	7	CDKN1A	8	GMNN	13
GAS2L1	7	DDB2	8	GRSF1	13
GINS3	7	DDX27	8	HACD3	13
IREB2	7	GMNN	8	IDH3A	13
KIF23	7	KIF18A	8	MAD2L1	13
KIF2C	7	KIF18B	8	MOB1A	13
MCM2	7	MCM2	8	MRPL37	13
MDM2	7	MDM2	8	MRPS7	13
NCAPH	7	MKI67	8	NAA50	13
NDC1	7	NCAPH	8	NDUFS3	13
NDE1	7	NDC1	8	NME1	13
ORC1	7	PAICS	8	RPL26L1	13
PAICS	7	PARPBP	8	SAE1	13
PBX1	7	PGAM1	8	SNRPF	13
PRC1	7	POLD2	8	TMEM126A	13
SERINC3	7	RFC5	8	TMEM167A	13
SMAD4	7	RNF19B	8	TMPO	13
SMCHD1	7	SNRPF	8	TRA2B	13
SNRPD1	7	STIL	8	UBE2S	13

GSE13294		**GSE17536**		**GSE26682**	
Gene Name	Commonalities for all pathway members	Gene Name	Commonalities for all pathway members	Gene Name	Commonalities for all pathway members
SPATA18	7	SUV39H2	8	DHX36	13
SUV39H2	7	TIPIN	8	ICOSLG	13
TIPIN	7	TSPAN6	8	LLGL1	13
TMEM167A	7	AAR2	7	SF3B5	13
TNFSF9	7	ARHGAP11A	7	TMED2	13
TNNC2	7	AURKA	7	TMEM70	13
TSR1	7	C18orf21	7	UTP18	13
UBE2S	7	CCDC88A	7	ZNHIT3	13
VAPA	7	CDCA3	7	HAUS1	13
ZCCHC2	7	CDK2	7	ARL1	13
ZWINT	7	CHEK1	7	CAND1	13
ABCE1	6	CKS2	7	MMADHC	13
AEN	6	DCUN1D5	7	OSTC	13
ARID3A	6	DEPDC1	7	PBK	13
ARL6IP5	6	DLGAP5	7	NCBP1	13
ASCL2	6	DTYMK	7	ACAT1	12
ASF1B	6	EXOSC3	7	AIFM1	12
ASPHD2	6	EZH2	7	CACYBP	12
ATP5B	6	FANCI	7	CHEK1	12
BUB1	6	HMBS	7	EMC8	12
C14orf142	6	HSPE1	7	FANCI	12
C1QBP	6	INO80C	7	GLRX3	12
C5orf15	6	INPP1	7	GRPEL1	12
CBX5	6	KIF15	7	ILF2	12
CCNB2	6	KIF2C	7	MRPS30	12
CCT2	6	MTFP1	7	PAICS	12

GSE13294		GSE17536		GSE26682	
Gene Name	Commonalities for all pathway members	Gene Name	Commonalities for all pathway members	Gene Name	Commonalities for all pathway members
CD274	6	MTHFD2	7	POLE3	12
CDCA2	6	NEK2	7	RFC5	12
CDCA8	6	NME1	7	RRM2	12
CDKN3	6	NUF2	7	SDHB	12
CENPM	6	NUP37	7	SLBP	12
CHD6	6	NUSAP1	7	TIMM17A	12
CHEK1	6	PA2G4	7	TSN	12
CLPX	6	PLK4	7	AFF4	12
CRK	6	RHOF	7	AP5M1	12
CSNK1A1	6	RPS27L	7	C22orf42	12
CTPS1	6	RRM2	7	SUZ12	12
DCUN1D5	6	SARNP	7	ZBED5	12
DEPDC1	6	SNRPG	7	WDR75	12
DEPDC1B	6	SPC25	7	CNPY2	12
DNMT1	6	TMPO	7	EIF4E	12
EEF1E1	6	TTK	7	FAM96A	12
EXO1	6	UBE2T	7	G3BP2	12
FAAP100	6	UNG	7	GCSH	12
FARP1	6	YTHDF1	7	KIF2A	12
FERMT2	6	ANXA2P2	6	MRPL3	12
GLRX3	6	ARMCX1	6	SCO1	12
GRPEL1	6	ASPM	6	VPS4B	12
GSE1	6	ATP9A	6	EIF3J	12
HAT1	6	BAG3	6	LOC101929280	12
IDH3A	6	C18orf25	6	SNRPG	12
IRF2BP2	6	C5orf15	6	DDX50	12

GSE13294		GSE17536		GSE26682	
Gene Name	Commonalities for all pathway members	Gene Name	Commonalities for all pathway members	Gene Name	Commonalities for all pathway members
ISG20	6	CACNA2D1	6	METAP2	12
JAK2	6	CCDC109B	6	PRPF18	12
JKAMP	6	CCDC68	6	RBM7	12
KIF4A	6	CDC25A	6	GPR137	12
KLHL28	6	CDC25C	6	ACLY	11
KNSTRN	6	CDCA8	6	AURKB	11
LAP3	6	CDKN3	6	BIRC5	11
LEO1	6	CENPK	6	C16orf59	11
LOC101929280	6	CENPM	6	CCT7	11
MAPK1	6	CIB1	6	CDCA5	11
MBD2	6	CKS1B	6	CKS2	11
MCM4	6	COLEC12	6	DNAJC9	11
MELK	6	CYSTM1	6	DTYMK	11
MTFP1	6	DBF4	6	FEN1	11
MTFR2	6	DENND5A	6	FXN	11
MTHFD2	6	DNAJB4	6	HSPE1	11
NAA38	6	DNMT1	6	IMMT	11
NCAPD3	6	E2F8	6	LSM2	11
NCAPG	6	ELK3	6	MCM2	11
NCAPG2	6	EPHA2	6	MRPL21	11
NELFCD	6	EXO1	6	NDUFAF4	11
NME1	6	FDXR	6	PNP	11
NUDT5	6	FOXM1	6	POLR1B	11
ODF3B	6	GINS1	6	POLR3K	11
OIP5	6	GPR160	6	PRPF38A	11
PARPBP	6	HNRNPL	6	PSMD9	11

GSE13294		**GSE17536**		**GSE26682**	
Gene Name	Commonalities for all pathway members	Gene Name	Commonalities for all pathway members	Gene Name	Commonalities for all pathway members
PCM1	6	IFT52	6	RFC2	11
PLK4	6	KIF20A	6	RFC4	11
POFUT1	6	KIF4A	6	RNASEH2A	11
POLQ	6	KNSTRN	6	SHMT2	11
PPP2R2A	6	KPNA2	6	SLC35B1	11
PSMD9	6	KRT8	6	SMC2	11
PSME1	6	LYAR	6	SNRNP25	11
RASGRP1	6	MAFF	6	SSRP1	11
RCC1	6	MCM4	6	TIMELESS	11
RFC1	6	MIS18A	6	TK1	11
RFC5	6	MRPL19	6	TUBA1C	11
RNF138	6	NCAPG	6	UBE2N	11
SCO1	6	NOP58	6	UBE2T	11
SET	6	NRP1	6	VDAC1	11
SHANK2	6	OLFML2B	6	C14orf178	11
SHROOM4	6	PGP	6	MON2	11
SLBP	6	PHLDA2	6	NEUROD4	11
SLC35C2	6	PMP22	6	OXT	11
SNRPF	6	PNPT1	6	PSMC6	11
SOCS6	6	POFUT1	6	SMARCA5	11
SPAG5	6	POLE2	6	TMEM145	11
SRP72	6	PPP2R2A	6	LOC220077	11
SYPL1	6	PSMD9	6	MUC8	11
TCTN3	6	PTRF	6	COPZ1	11
TIMM21	6	QKI	6	DBI	11
TM7SF3	6	RAB27B	6	MRPL15	11

GSE13294		GSE17536		GSE26682	
Gene Name	Commonalities for all pathway members	Gene Name	Commonalities for all pathway members	Gene Name	Commonalities for all pathway members
TNFRSF10D	6	RFC2	6	MRPS28	11
TRIM22	6	RNASEH2A	6	MTX2	11
TRIP13	6	SDHB	6	SPCS1	11
TYMS	6	SFRP2	6	WDR61	11
UBE2T	6	SGCB	6	PIK3C3	11
VPS4B	6	SKA1	6	TIMM21	11
WARS	6	SNRPA	6	BID	11
YME1L1	6	SRA1	6	CCT2	11
YWHAB	6	TACC3	6	DDX46	11
ACTR6	5	TAF4	6	ETFA	11
ADAMTS1	5	TOP2A	6	HNRNPLL	11
ADPRHL2	5	TPX2	6	MAP2K1	11
AIDA	5	TSPYL5	6	MRPS23	11
AIFM3	5	VGLL3	6	NDUFA11	11
AMD1	5	WDHD1	6	NUDCD2	11
ANLN	5	ZMAT3	6	PAIP2	11
ANTXR2	5	ZWILCH	6	PRKRA	11
ARHGAP30	5	ANTXR1	5	RAB1A	11
ASPM	5	ANXA1	5	SELT	11
ATP6V1B2	5	ASCL2	5	TMX1	11
BCL2L12	5	ASF1B	5	TRAM1	11
BRCA1	5	BCAT1	5	TYMS	11
BUD13	5	BNIP3L	5	ATP6V1B2	11
CAPRIN1	5	BTG3	5	BCAS2	11
CASC5	5	C19orf33	5	CHMP2B	11
CCDC109B	5	C2orf47	5	DYM	11

GSE13294		GSE17536		GSE26682	
Gene Name	Commonalities for all pathway members	Gene Name	Commonalities for all pathway members	Gene Name	Commonalities for all pathway members
CCL4	5	CALD1	5	NRBF2	11
CCNDBP1	5	CCNE2	5	SCYL2	11
CDC123	5	CCNF	5	GSK3A	11
CDC42EP1	5	CDC42SE1	5	DHX9	11
CDC6	5	CDCA4	5	RPL36AL	11
CDH1	5	CDK14	5	RNF19B	11
CDK1	5	CDX2	5	AIMP2	10
CDK2	5	CENPN	5	ARMC10	10
CHAC2	5	CLIC4	5	CANX	10
CHAF1B	5	COL8A1	5	CDC123	10
CHIC2	5	COQ10B	5	CDC6	10
CKS1B	5	CPNE1	5	CDKN3	10
CLEC2B	5	CTGF	5	CENPN	10
CLN6	5	DGCR6L	5	CFAP20	10
CNPY2	5	DMWD	5	CHCHD3	10
CNRIP1	5	DNA2	5	CHCHD6	10
COPS4	5	DOCK5	5	COPRS	10
CTPS2	5	ECE2	5	CORO1C	10
CYB5D1	5	ECM2	5	CYC1	10
CYB5D2	5	EIF4G1	5	GADD45GIP1	10
DBF4	5	EMC7	5	GGCX	10
DDB2	5	ERCC6L	5	GPATCH4	10
DDR2	5	ESPL1	5	GSTCD	10
DLGAP5	5	EXOSC9	5	HCCS	10
DSCR3	5	FAM46A	5	HNRNPD	10
DYNC1LI2	5	FANCD2	5	LSM4	10

GSE13294		GSE17536		GSE26682	
Gene Name	Commonalities for all pathway members	Gene Name	Commonalities for all pathway members	Gene Name	Commonalities for all pathway members
EBF1	5	FECH	5	MAPRE1	10
ECE2	5	FERMT2	5	MRPL35	10
EFEMP1	5	FLII	5	MRPS17	10
EFEMP2	5	FSTL1	5	NOLC1	10
EIF4E	5	GBP2	5	NUP205	10
ELAC2	5	GLIS2	5	OIP5	10
ERI1	5	GNB4	5	PPIH	10
ETFA	5	GSKIP	5	PSMB3	10
FAS	5	GTF2A2	5	PSMC3	10
FBXO22	5	H2AFX	5	PSMD12	10
FBXO45	5	HEG1	5	RACGAP1	10
FMO4	5	HELLS	5	RBM8A	10
GCNT7	5	HEXIM1	5	SKA1	10
GFM2	5	HJURP	5	SLIRP	10
GGT7	5	HPSE	5	SSR1	10
GLUD2	5	HSPA4L	5	SUV39H2	10
GNG11	5	ITGA5	5	TUBB	10
GPC6	5	JAM3	5	CSN3	10
GPD1L	5	KCNE4	5	DCST1	10
GPR160	5	KDSR	5	FAR1	10
GRPEL2	5	KNTC1	5	GCH1	10
GRSF1	5	LAMB2	5	GNPTAB	10
GSKIP	5	LAMB3	5	GPD2	10
GTF2B	5	LHFP	5	IBTK	10
GZMA	5	LMCD1	5	MYOC	10
H2AFZ	5	LOC340107	5	NUFIP2	10

GSE13294		GSE17536		GSE26682	
Gene Name	Commonalities for all pathway members	Gene Name	Commonalities for all pathway members	Gene Name	Commonalities for all pathway members
HBS1L	5	MAP1B	5	RAB10	10
HSPE1	5	MCAT	5	RAB6B	10
IER3IP1	5	MCM3	5	SLC5A2	10
IFIT3	5	MCMBP	5	TROVE2	10
IHH	5	ME2	5	TVP23B	10
IL1R1	5	MED31	5	USO1	10
ITGA5	5	MFAP5	5	EIF5B	10
KCTD9	5	MIF	5	NIFK	10
KDSR	5	MIR100HG	5	ATP4B	10
KIF14	5	MPHOSPH9	5	DTX2P1UPK3BP1-PMS2P11	10
KIF18A	5	MRPL17	5	ADNP2	10
KIF18B	5	MRPL44	5	AMD1	10
LINC00543	5	NCAPD3	5	C14orf166	10
LINC00672	5	NCAPG2	5	CDK4	10
LMNB2	5	NCOA6	5	COPS3	10
LYSMD2	5	NDEL1	5	CYB5A	10
MAP3K6	5	NELFCD	5	DCK	10
MBP	5	NMT1	5	ELAC1	10
MCM3	5	NNMT	5	ETF1	10
MCMBP	5	NOL4L	5	FECH	10
MDH1	5	NTM	5	MRPL30	10
MFAP1	5	NUTF2	5	MRPL45	10
MINPP1	5	OSER1	5	ORMDL2	10
MPHOSPH9	5	PDGFRL	5	P4HB	10
MRPL11	5	PDLIM7	5	PRDX4	10

GSE13294		GSE17536		GSE26682	
Gene Name	Commonalities for all pathway members	Gene Name	Commonalities for all pathway members	Gene Name	Commonalities for all pathway members
MRPL44	5	PDRG1	5	SLC25A5	10
MRPL46	5	PHLDA1	5	SNRPD2	10
MRPS11	5	PKD2	5	TM9SF1	10
MRPS12	5	PKP3	5	TMEM97	10
MTA2	5	PLAUR	5	TRIAP1	10
NASP	5	PLEK2	5	UBE2D2	10
NCLN	5	PLK3	5	ARHGEF6	10
NDC80	5	POLQ	5	HDHD2	10
NEIL2	5	POLR1C	5	RAB27B	10
NOP16	5	POLR3K	5	RNF138	10
NRBF2	5	PPIB	5	SOCS6	10
NUP37	5	PPIL1	5	TNFAIP8	10
ORC6	5	PRLR	5	ZCCHC2	10
PBK	5	PRRX1	5	ANAPC16	10
PIGU	5	PSMC4	5	CYB5R4	10
PIK3C3	5	PSMD14	5	DAD1	10
PLCB4	5	RAB31	5	DCP1A	10
PLK1	5	RAD54L	5	EIF4A3	10
PLK2	5	RANGAP1	5	GNAI3	10
PQLC1	5	RASSF8	5	HIAT1	10
PRIM1	5	RFC4	5	ISCA1	10
PRMT5	5	S100A14	5	LLPH	10
PRPF39	5	SERPINH1	5	LSM6	10
QPRT	5	SETBP1	5	MAD2L1BP	10
RABEP1	5	SFXN1	5	MFAP1	10
RAD51	5	SLC25A10	5	OST4	10

GSE13294		**GSE17536**		**GSE26682**	
Gene Name	Commonalities for all pathway members	Gene Name	Commonalities for all pathway members	Gene Name	Commonalities for all pathway members
RAD51AP1	5	SLC35C2	5	PCNP	10
RAP2A	5	SLF1	5	PDCD10	10
RNASEH2A	5	SPARC	5	PELO	10
RNF125	5	SRBD1	5	PGM2	10
RPL36AL	5	SSTR5	5	RARS	10
RPS27L	5	SYT7	5	RPF1	10
RRM2	5	TARBP2	5	SRP72	10
RTCA	5	THBS2	5	TMEM14B	10
RUVBL1	5	TIMM21	5	TPRKB	10
SELT	5	TMEM30A	5	VDAC3	10
SFXN1	5	TNFRSF10B	5	YWHAQ	10
SHMT2	5	TNFSF13	5	AASDHPPT	10
SLC11A1	5	TP53RK	5	ACTR10	10
SLC25A37	5	TRIM22	5	ARMT1	10
SLK	5	TTF2	5	C14orf119	10
SNRPG	5	TTLL12	5	CUL5	10
SOX4	5	TWIST1	5	PPTC7	10
SP100	5	TYMP	5	C9orf117	10
SPDL1	5	TYMS	5	INSL3	10
SS18	5	UBE2C	5	LDLRAD4	10
THBS2	5	VCAN	5	LRRC43	10
TK1	5	VEGFC	5	PRDM12	10
TNFAIP8	5	VPS37B	5	NIF3L1	10
TOP2A	5	WDR12	5	NOL11	10
TPX2	5	WDR41	5	MCL1	10
TRA2B	5	XRCC5	5	TBC1D15	10

GSE13294		GSE17536		GSE26682	
Gene Name	Commonalities for all pathway members	Gene Name	Commonalities for all pathway members	Gene Name	Commonalities for all pathway members
TTK	5	YEATS4	5	AP2S1	9
UHRF1	5	YWHAB	5	ASF1B	9
USP7	5	ZEB2	5	ATG101	9
37500	4	ZFPM2	5	BUB1B	9
AAR2	4	ZMYND8	5	CDK2	9
ABLIM3	4	38412	4	CKAP5	9
ACADVL	4	ABAT	4	CKS1B	9
ACSS2	4	ACACA	4	DDX19A	9
ADAM10	4	ACTL6A	4	DLGAP5	9
ADAM9	4	ADAM12	4	DNMT1	9
ADGRL4	4	ADNP2	4	EXO1	9
AFG3L2	4	ADRM1	4	EXOSC3	9
AGPAT5	4	AEBP1	4	FDX1	9
AK6	4	AEN	4	GAPDH	9
AKT3	4	AFG3L2	4	GNB1	9
ALAS1	4	AKAP8	4	GPN3	9
ANGPTL2	4	AKT3	4	H2AFX	9
ANXA2P2	4	ANKRD10-IT1	4	KIF11	9
AP3M1	4	ANKRD27	4	LMNB2	9
ARF3	4	ANXA2P1	4	MCM10	9
ARHGAP11A	4	ANXA2P3	4	MCM7	9
ARHGAP9	4	AP1S3	4	MED17	9
ARPC5	4	APOL1	4	MRPL17	9
ARPP19	4	ARFGAP1	4	MRPL19	9
ATAD2	4	ARFGEF2	4	NAP1L4	9
ATP6V0D1	4	ARRDC3	4	NCAPD3	9

GSE13294		GSE17536		GSE26682	
Gene Name	Commonalities for all pathway members	Gene Name	Commonalities for all pathway members	Gene Name	Commonalities for all pathway members
ATP6V1A	4	ATF1	4	NUDT5	9
AUNIP	4	ATP5D	4	PDCD5	9
BAG3	4	BASP1	4	PDSS1	9
BAK1	4	BCAS2	4	PIGX	9
BARHL1	4	BCL11A	4	PNO1	9
BASP1	4	BCL6	4	PPP1R8	9
BCCIP	4	BGN	4	PRMT5	9
BLM	4	BICC1	4	PSMD14	9
BRI3BP	4	BID	4	RBM28	9
C11orf96	4	BLVRA	4	RRM1	9
C18orf21	4	BRCA1	4	SMNDC1	9
C1D	4	C10orf10	4	TALDO1	9
C1R	4	C10orf2	4	TPX2	9
C20orf194	4	C1R	4	TRMT112	9
C8orf58	4	CALU	4	TSR1	9
CABIN1	4	CAND1	4	UMPS	9
CAV1	4	CAPN1	4	WDR12	9
CCAR2	4	CASP4	4	YIF1A	9
CCDC154	4	CBFA2T2	4	ARL6IP1	9
CCNA2	4	CCAR2	4	B3GNT4	9
CCNE2	4	CCDC27	4	C2orf16	9
CCNF	4	CCT7	4	HPYR1	9
CD93	4	CDC7	4	KCNN1	9
CDC20	4	CDCP1	4	MBD4	9
CDC25C	4	CDH11	4	NFIA-AS2	9
CDC40	4	CDK13	4	PHF3	9

GSE13294		GSE17536		GSE26682	
Gene Name	Commonalities for all pathway members	Gene Name	Commonalities for all pathway members	Gene Name	Commonalities for all pathway members
CDCP1	4	CECR5	4	TMX3	9
CDT1	4	CENPA	4	VGLL1	9
CECR5	4	CENPF	4	VPS13C	9
CENPA	4	CEP164	4	ZNF322	9
CENPH	4	CERCAM	4	ZNF645	9
CENPU	4	CFAP36	4	HTR1B	9
CENPW	4	CHAF1A	4	DDX18	9
CHCHD10	4	CLIP4	4	LRRC40	9
CHMP2B	4	CLTB	4	MAP3K2	9
CHMP7	4	CMTM3	4	OR7E104P	9
CHN2	4	COL10A1	4	AGPS	9
CHRNB2	4	COL11A1	4	ANKFY1	9
CHST11	4	COL5A1	4	ATP6V0B	9
CITED2	4	COL5A2	4	EBP	9
CLINT1	4	CPOX	4	FDXR	9
CMSS1	4	CSE1L	4	GORASP2	9
COL18A1	4	CSGALNACT2	4	HAUS6	9
COL3A1	4	CXCR5	4	IMP3	9
COL6A2	4	CYB5D1	4	MANF	9
COPS3	4	CYBRD1	4	ME2	9
COQ10A	4	CYP1B1	4	PLGRKT	9
COX8A	4	CYR61	4	RBX1	9
CPNE1	4	DCTN6	4	SLC25A11	9
CRCP	4	DDR2	4	TMX2	9
CTDNEP1	4	DDX39A	4	TOMM22	9
CYR61	4	DDX54	4	UQCRC1	9

GSE13294		**GSE17536**		**GSE26682**	
Gene Name	Commonalities for all pathway members	Gene Name	Commonalities for all pathway members	Gene Name	Commonalities for all pathway members
DCK	4	DEFB124	4	C18orf25	9
DCN	4	DERL2	4	CCDC68	9
DCP1A	4	DHFR	4	DDB2	9
DDX59	4	DHX15	4	GSKIP	9
DHFR	4	DIABLO	4	MDM2	9
DIAPH3	4	DIDO1	4	MED31	9
DKFZP586I1420	4	DNAJC5	4	NAA38	9
DOCK5	4	DOCK4	4	NPTN	9
DONSON	4	DPM1	4	PAFAH1B1	9
DSC2	4	DPYSL3	4	RNASE4	9
DSCC1	4	DSCC1	4	RPS27L	9
DTL	4	DUSP5	4	SMAD2	9
DUSP18	4	ELMO2	4	TSNARE1	9
E2F8	4	EMC8	4	UBE2G1	9
ECT2	4	ENO2	4	ABCB10	9
EHD2	4	EP300-AS1	4	ADH5	9
EIF4A3	4	ERO1A	4	ANAPC10	9
ELP2	4	ERRFI1	4	AP3M1	9
ENTPD7	4	F3	4	ATP6V1C1	9
EPG5	4	FANCA	4	C14orf142	9
EPT1	4	FBL	4	C1D	9
ERCC6L	4	FBN1	4	COPS4	9
ERLEC1	4	FBXL7	4	CTDSPL2	9
ERO1A	4	FBXO22	4	DCTPP1	9
EVL	4	FCHO2	4	DNAJA1	9
EXOSC2	4	FLT4	4	GDI2	9

GSE13294		GSE17536		GSE26682	
Gene Name	Commonalities for all pathway members	Gene Name	Commonalities for all pathway members	Gene Name	Commonalities for all pathway members
EXOSC3	4	FLVCR1	4	GTF2A2	9
FAM101B	4	FSCN1	4	IARS	9
FAM169A	4	GADD45B	4	IREB2	9
FAM96A	4	GAPDH	4	KIF23	9
FBN1	4	GAS1	4	LAP3	9
FBXL7	4	GAS2L1	4	LSM5	9
FLII	4	GCN1	4	MMGT1	9
FRMD6	4	GCNT3	4	NFIL3	9
FXN	4	GCSH	4	PGRMC1	9
GAS1	4	GGCT	4	RFC3	9
GCSH	4	GGT5	4	TCTN3	9
GGA1	4	GID8	4	VBP1	9
GINS2	4	GINS2	4	ACTR6	9
GJA1	4	GJA1	4	FBXO8	9
GMEB2	4	GJB3	4	ISOC1	9
GNAI3	4	GMPS	4	ORC4	9
GPN2	4	GOSR2	4	PANK3	9
GTF2A2	4	GPC6	4	SLK	9
GTF2E2	4	GPR20	4	SRP54	9
GYS1	4	GTF2B	4	ZBTB1	9
GZMH	4	H2AFY	4	ASXL2	9
HACD3	4	HAT1	4	CRYAB	9
HMBS	4	HAUS3	4	GRINA	9
HNRNPL	4	HIST2H2AA3	4	MARK2	9
HSPA14	4	HMMR	4	PDLIM7	9
ICAM1	4	HOOK1	4	WHAMMP2	9

GSE13294		**GSE17536**		**GSE26682**	
Gene Name	Commonalities for all pathway members	Gene Name	Commonalities for all pathway members	Gene Name	Commonalities for all pathway members
ICMT	4	HOXD12	4	SLC22A14	9
IFIT5	4	HTRA1	4	SLC5A5	9
IFRD2	4	IER3IP1	4	MIS18A	9
IL10RA	4	IER5	4	PRIM1	9
IL1RN	4	IGF2-AS	4	HNRNPU	9
INO80C	4	IGH	4	SH3GLB1	9
ITGAL	4	IL1R1	4	CBX1	9
ITGB1	4	INHBA	4	CYP2A7P1	9
ITK	4	ISCU	4	AURKA	8
JAM3	4	ITGA3	4	CCDC34	8
KCNK5	4	JAK2	4	CCNB2	8
KIF11	4	JAM2	4	CDC20	8
KIF1BP	4	KCNH5	4	CDC45	8
KIF2A	4	KCNT1	4	CDCA8	8
LFNG	4	KCTD9	4	CEP55	8
LGALS1	4	KIAA1462	4	CHAF1B	8
LINC01560	4	KIF1C	4	CIAPIN1	8
LMBR1	4	KLF14	4	CMSS1	8
LOC101928881	4	KLHDC1	4	CNIH4	8
LOC102725022	4	KLK10	4	DCAF10	8
LOC158402	4	KLK13	4	ECE2	8
LOC643733	4	KPNA3	4	FOXM1	8
LOX	4	LGALS1	4	GINS2	8
LOXL2	4	LGALS3	4	GPI	8
LRR1	4	LIF	4	ICT1	8
LRRC42	4	LIG1	4	KIF1BP	8

GSE13294		GSE17536		GSE26682	
Gene Name	Commonalities for all pathway members	Gene Name	Commonalities for all pathway members	Gene Name	Commonalities for all pathway members
LTV1	4	LINC00937	4	KIF20A	8
LYSMD3	4	LMNA	4	KIF2C	8
MAD2L1	4	LMNB2	4	MCM3	8
MAF	4	LMO7	4	MELK	8
MAP2K4	4	LOC101927040	4	MKI67	8
MAPRE1	4	LOC101928845	4	MND1	8
MCL1	4	LOX	4	MRPL14	8
MED11	4	LOXL1	4	MRPS2	8
MED13L	4	LRCH1	4	MTHFD2	8
MED17	4	LRRC42	4	NCAPG2	8
METAP2	4	LSM4	4	NCAPH	8
MEX3A	4	MAG	4	NDUFB9	8
MGP	4	MALT1	4	PARPBP	8
MIR34A	4	MBD2	4	PCBD1	8
MMADHC	4	MCFD2	4	PFKP	8
MOV10	4	MCM5	4	PICALM	8
MPDZ	4	MCM9	4		
MRPS15	4	MPDZ	4		
MRPS18C	4	MRPL14	4		
MRPS30	4	MRPL35	4		
NAA50	4	MRPL37	4		
NARS2	4	MRPS12	4		
NDUFAB1	4	MRPS15	4		
NDUFB5	4	MTA2	4		
NDUFS4	4	MTIF2	4		
NDUFV2	4	NAP1L3	4		

GSE13294		**GSE17536**		**GSE26682**	
Gene Name	Commonalities for all pathway members	Gene Name	Commonalities for all pathway members	Gene Name	Commonalities for all pathway members
NEK2	4	NAPG	4		
NOLC1	4	NCLN	4		
NUDCD2	4	NDN	4		
NUTM2B	4	NDUFV2	4		
PAIP2	4	NEDD1	4		
PDS5A	4	NIP7	4		
PES1	4	NOL10	4		
PFDN2	4	NOP10	4		
PHF23	4	NOTCH3	4		
PLA2G16	4	NOX4	4		
PLAUR	4	NR3C1	4		
POLR2A	4	NRBF2	4		
POLR3D	4	NUAK1	4		
PPIL1	4	NUDCD2	4		
PPP4R1	4	NUFIP1	4		
PRDX1	4	NUP107	4		
PRKAR1A	4	NUP50	4		
PRKAR2A	4	OGFOD1	4		
PRPF6	4	PCIF1	4		
PRRC1	4	PDGFRB	4		
PSMC3	4	PDLIM3	4		
PSMD8	4	PES1	4		
PTGES3	4	PFKP	4		
PXDN	4	PHF23	4		
RAB1A	4	PIK3CD-AS1	4		
RAB31	4	PLAGL2	4		

GSE13294		**GSE17536**		**GSE26682**	
Gene Name	Commonalities for all pathway members	Gene Name	Commonalities for all pathway members	Gene Name	Commonalities for all pathway members
RACGAP1	4	PLAU	4		
RALBP1	4	PLLP	4		
RANGAP1	4	PLOD1	4		
RARRES3	4	PLXNB2	4		
RARS	4	POLR1B	4		
RBM7	4	POLR2D	4		
REV3L	4	PPDPF	4		
RFC2	4	PPID	4		
RHBDD3	4	PPIF	4		
RNMTL1	4	PPP1CA	4		
RPA1	4	PPP1R3D	4		
RPIA	4	PRKD1	4		
RPL41	4	PRPF19	4		
RTFDC1	4	PSMA1	4		
RTTN	4	PSMA4	4		
S100A14	4	PSMA7	4		

Commonality Tables for Data Used in Chapter 4 (B)

GSE14333-USA		GSE14333-Melbourn		GSE4183-Normal colon	
Gene Name	Commonalities for all pathway members	Gene Name	Commonalities for all pathway members	Gene Name	Commonalities for all pathway members
MND1	13	CEP55	14	SHMT2	13
CEP55	11	SHMT2	14	MCM3	11
FEN1	11	BIRC5	12	TIMELESS	10
RRM1	11	CDC45	12	DCUN1D5	9
BUB1	10	CDK1	12	FANCI	9
DNAJC9	10	CCNA2	11	TRIP13	9
MTHFD2	10	CDC6	11	UBE2S	9
CCNB1	9	CENPM	11	BUB1	8
CCNB2	9	KIF2C	11	CDCA5	8
CDC45	9	KPNA2	11	MCM2	8
CDC6	9	NCAPG	11	MND1	8
CDCA5	9	RRM2	11	RACGAP1	8
CKS1B	9	SNRPF	11	RFC4	8
MCM10	9	TACC3	11	RFC5	8
NDC1	9	BRIP1	10	SUV39H2	8
NME1	9	BUB1	10	TIPIN	8
OIP5	9	BUB1B	10	TNFRSF10B	8
PBK	9	C16orf59	10	BIRC5	7
POLR3K	9	C1QBP	10	BUB1B	7

GSE14333-USA		**GSE14333-Melbourn**		**GSE4183-Normal colon**	
Gene Name	Commonalities for all pathway members	Gene Name	Commonalities for all pathway members	Gene Name	Commonalities for all pathway members
RAD51AP1	9	CDC20	10	CCT7	7
TIPIN	9	DNAJC9	10	CDK1	7
ANLN	8	FEN1	10	CKS1B	7
BIRC5	8	MAD2L1	10	KIF2C	7
CCNA2	8	MCM2	10	LMNB2	7
CDC20	8	NUP37	10	MCM10	7
CDKN3	8	RACGAP1	10	NME1	7
CHEK1	8	RAN	10	NUP37	7
DLGAP5	8	RFC5	10	PAICS	7
ECT2	8	SNRPD1	10	PRIM1	7
FANCI	8	SPAG5	10	RNASEH2A	7
KIAA1462	8	TNFSF9	10	SLC39A6	7
KIF14	8	UBE2S	10	SNRPD1	7
KIF18A	8	ANLN	9	SNRPF	7
KIF23	8	CCNB1	9	TPX2	7
KPNA2	8	CDCA5	9	UBE2T	7
MAD2L1	8	DEPDC1	9	ANLN	6
MKI67	8	DLGAP5	9	AURKB	6
MRPL44	8	DOCK5	9	BRCA1	6
NEK2	8	ELAVL1	9	CDC6	6
PAICS	8	INO80C	9	CDCA3	6
PGAM1	8	KIF23	9	DBF4	6
RACGAP1	8	MCAT	9	ECT2	6

GSE14333-USA		GSE14333-Melbourn		GSE4183-Normal colon	
Gene Name	Commonalities for all pathway members	Gene Name	Commonalities for all pathway members	Gene Name	Commonalities for all pathway members
RFC5	8	MND1	9	EEF1E1	6
SHMT2	8	NCAPH	9	INHBA	6
SUV39H2	8	NOP16	9	KIF23	6
TNFSF9	8	PA2G4	9	MAD2L1	6
TRIP13	8	PBK	9	MCM4	6
TTK	8	PLK1	9	MELK	6
UBE2S	8	PLK2	9	NCAPG2	6
UNG	8	RNASEH2A	9	NEK2	6
ZWINT	8	SNRPA1	9	NUSAP1	6
AURKB	7	TYMS	9	ORC6	6
C18orf25	7	ZWINT	9	PBK	6
CASC5	7	ADNP2	8	RAD51AP1	6
CDC25A	7	ANXA2P2	8	RAN	6
CDCA3	7	AP1S3	8	RFC2	6
CDK1	7	AURKB	8	CCNA2	5
CENPK	7	C18orf25	8	CDC45	5
CKS2	7	CBFA2T2	8	CDKN3	5
DDX27	7	CDCA3	8	CEP55	5
DNAJB4	7	CIB1	8	CHEK1	5
JAM3	7	CKS2	8	DDB2	5
KIF11	7	DBF4	8	DLGAP5	5
KIF18B	7	DCUN1D5	8	DNAJC9	5
KIF2C	7	DNMT1	8	DTL	5
MCAT	7	DTYMK	8	DUSP4	5
MDM2	7	EXO1	8	H2AFZ	5
MELK	7	EXOSC3	8	KIF4A	5

GSE14333-USA		GSE14333-Melbourn		GSE4183-Normal colon	
Gene Name	Commonalities for all pathway members	Gene Name	Commonalities for all pathway members	Gene Name	Commonalities for all pathway members
MTFP1	7	FOXM1	8	MKI67	5
ORC1	7	KIF11	8	MTHFD2	5
PA2G4	7	KIF4A	8	POLE2	5
PARPBP	7	LAMB3	8	PRC1	5
PGP	7	MCM4	8	RRM2	5
PHF20	7	MDM2	8	SMC4	5
PLK4	7	MELK	8	TOP2A	5
PRC1	7	MRPS12	8	UTP18	5
PTRF	7	NDC1	8	VRK1	5
RNF19B	7	NEK2	8	ASPM	4
SFXN1	7	NUF2	8	CASC5	4
SPC25	7	PAICS	8	CCNB1	4
STIL	7	PARPBP	8	CCNB2	4
TIMELESS	7	PRC1	8	CDK2	4
TIMM13	7	PSMD14	8	CKS2	4
TK1	7	PSMD9	8	DNMT1	4
TMPO	7	RAD51AP1	8	FDXR	4
TOP2A	7	RPS27L	8	FEN1	4
TSPAN6	7	SOCS6	8	GINS2	4
WDHD1	7	TK1	8	HAT1	4
AAR2	6	TMPO	8	KIF11	4
ACTL6A	6	TNFRSF10A	8	KIF20A	4
AKT3	6	ZEB2	8	MAF	4
AP1S3	6	AEN	7	NCAPH	4
ARHGAP11A	6	ASPM	7	ORC1	4
ARL4C	6	CACNG6	7	PARPBP	4

GSE14333-USA		GSE14333-Melbourn		GSE4183-Normal colon	
Gene Name	Commonalities for all pathway members	Gene Name	Commonalities for all pathway members	Gene Name	Commonalities for all pathway members
AURKA	6	CACYBP	7	PLK4	4
BRCA1	6	CASC5	7	TK1	4
BUB1B	6	CCNB2	7	WISP1	4
C2orf47	6	CCT3	7	ZWINT	4
CAND1	6	CDCA8	7	ANXA2P2	3
CASQ2	6	CDCP1	7	C1QBP	3
CDCA4	6	CHEK1	7	CDC25A	3
CDCA8	6	COPS3	7	CDC25C	3
CDKN1A	6	CPSF6	7	CDCA4	3
CENPA	6	DDX27	7	CDCA8	3
CENPM	6	DUSP4	7	CENPN	3
CERCAM	6	ECT2	7	DEPDC1	3
DEPDC1	6	FANCI	7	DSCC1	3
DNA2	6	FBXO45	7	EXOSC3	3
EEF1E1	6	FDXR	7	HSPA4L	3
EMC8	6	FERMT2	7	KCTD9	3
EPHA2	6	GINS2	7	KIF18A	3
EXOSC3	6	GMPS	7	MDM2	3
FAM46A	6	H2AFZ	7	NCAPD3	3
FBXL7	6	INPP1	7	NUDT5	3
FDXR	6	ITGB4	7	OIP5	3
GINS1	6	KDSR	7	PTTG1	3
GINS2	6	KIF18B	7	SHCBP1	3
GMNN	6	KIF20A	7	SLC25A4	3
H2AFZ	6	KITLG	7	SMC2	3
HNRNPL	6	LAD1	7	TAF4	3

GSE14333-USA		GSE14333-Melbourn		GSE4183-Normal colon	
Gene Name	Commonalities for all pathway members	Gene Name	Commonalities for all pathway members	Gene Name	Commonalities for all pathway members
HSPE1	6	LIF	7	TUBB6	3
INO80C	6	LIG1	7	TYMS	3
ITGA5	6	MAGOHB	7	ZMAT3	3
KIF20B	6	MBD2	7	NUP107	10
LGALS1	6	MCM3	7	MCM7	9
MAP1B	6	MCM5	7	FOXM1	8
MCM2	6	MKI67	7	GINS1	8
MCM5	6	MRTO4	7	KIF14	8
MIS18A	6	MTHFD2	7	KPNA2	8
MRPL19	6	MUC6	7	PHLDA1	8
NCAPH	6	NME1	7	SKP2	8
NDC80	6	NUSAP1	7	CDK4	7
NMT1	6	OIP5	7	EZH2	7
NRP1	6	PLIN3	7	GMPS	7
NUF2	6	POLD2	7	GTF2E2	7
OGFOD1	6	PRIM1	7	LIG1	7
OLFML2B	6	RAB27B	7	LYAR	7
PDLIM7	6	RBM28	7	NOP58	7
PHF5A	6	RFC2	7	NUF2	7
PMP22	6	RFC4	7	NUP205	7
POLD2	6	RRM1	7	POLD2	7
POLE2	6	SHCBP1	7	RUVBL1	7
POLR2D	6	SPATA5L1	7	S100A11	7
PRIM1	6	STIP1	7	UBE2C	7
RAB27B	6	TIMELESS	7	ASF1B	6
RAD51	6	TUBB	7	C1orf112	6

GSE14333-USA		GSE14333-Melbourn		GSE4183-Normal colon	
Gene Name	Commonalities for all pathway members	Gene Name	Commonalities for all pathway members	Gene Name	Commonalities for all pathway members
RAN	6	UBE2T	7	HJURP	6
RRM2	6	WDTC1	7	KIF18B	6
SHANK2	6	ZWILCH	7	NDC1	6
SLC25A39	6	ADRBK1	6	SNRPA	6
SNRPF	6	ALAS1	6	SPC25	6
TP53RK	6	ANKRD39	6	TSR1	6
TTPAL	6	ASF1B	6	ZWILCH	6
UBE2C	6	AURKA	6	BID	5
UQCC1	6	BRCA1	6	CENPA	5
VGLL3	6	CCDC68	6	CENPK	5
WDR12	6	CCT7	6	DTYMK	5
YTHDF1	6	CDX2	6	FBL	5
ZWILCH	6	CENPK	6	LATS2	5
ADAMTS2	5	CEP68	6	MIS18A	5
AEBP1	5	CKS1B	6	MRPL17	5
AGR2	5	CLIP4	6	MRPS17	5
ANGPTL1	5	CLPP	6	MRPS2	5
ANXA2P2	5	DDB2	6	MRPS23	5
ASPM	5	DDX54	6	NDC80	5
BCL6	5	DHX9	6	PA2G4	5
C11orf96	5	DSPP	6	PHLDA3	5
CACNA2D1	5	DTL	6	POLR2D	5
CCDC109B	5	E2F8	6	RRM1	5
CD93	5	EFEMP1	6	SSRP1	5
CDK2	5	EPHA2	6	TNFRSF10A	5
CHSY1	5	EPS8L1	6	TRAP1	5

GSE14333-USA		GSE14333-Melbourn		GSE4183-Normal colon	
Gene Name	Commonalities for all pathway members	Gene Name	Commonalities for all pathway members	Gene Name	Commonalities for all pathway members
CMC2	5	EWSR1	6	TTK	5
CMTM3	5	EXOSC2	6	ADGRG1	4
COL5A1	5	EZH2	6	ASCL2	4
COL5A2	5	F3	6	AURKA	4
COLEC12	5	FBXO22	6	C1D	4
DBF4	5	FLII	6	CKAP2	4
DCUN1D5	5	HAMP	6	DENND5A	4
DDB2	5	HELLS	6	EMC8	4
DDR2	5	HJURP	6	FANCD2	4
DDX39A	5	HNRNPU	6	HMMR	4
DGCR6L	5	IER3IP1	6	KIAA1524	4
DHFR	5	IER5	6	MCAT	4
DIDO1	5	KIF14	6	MRTO4	4
DNTTIP1	5	KIF18A	6	NASP	4
DTL	5	KNTC1	6	NCAPD2	4
E2F8	5	LHFP	6	NDN	4
ECE2	5	LINC00491	6	PGM2	4
EHD4	5	LOC101928718	6	PSMD14	4
EXO1	5	LOC728099	6	RBM28	4
FAS	5	LRRC25	6	SLC7A11	4
FECH	5	LRRC8A	6	TACC3	4
FERMT2	5	MCM10	6	THY1	4
FOXM1	5	MGP	6	TIE1	4
GAS1	5	MSH2	6	TIMM50	4
GPLD1	5	MTDH	6	TNFRSF12A	4
HEG1	5	NCAPG2	6	TRIAP1	4

GSE14333-USA		GSE14333-Melbourn		GSE4183-Normal colon	
Gene Name	Commonalities for all pathway members	Gene Name	Commonalities for all pathway members	Gene Name	Commonalities for all pathway members
HELLS	5	NSDHL	6	B3GALT6	9
HJURP	5	OBFC1	6	CCT3	9
HMBS	5	ORC1	6	CMSS1	9
HSPA4L	5	PAPPA	6	GPN3	9
HSPD1	5	PCM1	6	ILF2	9
IER3IP1	5	PGAM1	6	MSH6	9
IFT52	5	PHF23	6	NUP155	9
IL1R1	5	PHLDA2	6	PHF19	9
ILF2	5	PLK4	6	ASUN	8
INPP1	5	POLR3K	6	CEP78	8
ISLR	5	PSMB6	6	FIGNL1	8
KAT2B	5	RAD51	6	MTHFD1	8
KCNF1	5	REEP4	6	NFE2L3	8
KIF4A	5	RFT1	6	RIPK2	8
KNSTRN	5	RNF126	6	UHRF1	8
KRT3	5	RUVBL2	6	AJUBA	7
LAMA3	5	SHB	6	C12orf10	7
LAMB2	5	SKA1	6	CEMIP	7
LHFP	5	SLC7A11	6	CKAP5	7
LIN9	5	SNCA	6	CTPS1	7
LSM4	5	SRPRB	6	DDIT4	7
MAP7D1	5	STT3A	6	DNAJA3	7
MCM3	5	SUV39H2	6	DUSP14	7
MCM4	5	TIMM50	6	E2F7	7
MFAP5	5	TIPIN	6	IARS	7
MITF	5	TNFRSF10D	6	LONP1	7

GSE14333-USA		GSE14333-Melbourn		GSE4183-Normal colon	
Gene Name	Commonalities for all pathway members	Gene Name	Commonalities for all pathway members	Gene Name	Commonalities for all pathway members
MPHOSPH9	5	TOMM22	6	MRGBP	7
MRPL17	5	TSFM	6	NTMT1	7
MRPS34	5	TTK	6	PCNA	7
NAP1L3	5	TUBA4A	6	PRPF4	7
NCAPD3	5	UBE2M	6	RFC3	7
NIFK	5	WDR12	6	TMEM161A	7
NOP58	5	WDR34	6	TMEM97	7
NUDT5	5	YEATS4	6	ABCA8	6
NUSAP1	5	YWHAB	6	ABCC1	6
ORC6	5	AAR2	5	ACD	6
PCM1	5	ABAT	5	ACTR3B	6
PDRG1	5	ACAT2	5	ADAMDEC1	6
PES1	5	ADAMTS1	5	ADRM1	6
PLAGL2	5	AGR2	5	ATP11A	6
PLAUR	5	AHNAK2	5	C1orf216	6
PLK1	5	AKAP12	5	CCDC59	6
PLK2	5	ANKFY1	5	CCT6A	6
PLK3	5	ANXA2P1	5	CENPH	6
POFUT1	5	ANXA2P3	5	CNN2	6
POLR1B	5	AOX1	5	CPSF3	6
PPIL1	5	ARFGEF2	5	CSE1L	6
PRRX1	5	ARHGAP11A	5	E2F6	6
PSMD14	5	ATIC	5	FOXQ1	6
PSMD9	5	BCCIP	5	FTSJ2	6
PTGES3	5	BLOC1S2	5	GALK1	6
PTTG1	5	C12orf10	5	ILF3	6

GSE14333-USA		GSE14333-Melbourn		GSE4183-Normal colon	
Gene Name	Commonalities for all pathway members	Gene Name	Commonalities for all pathway members	Gene Name	Commonalities for all pathway members
PUS1	5	C17orf51	5	IMPDH1	6
QKI	5	CACNA2D1	5	IPO9	6
RAB31	5	CAND1	5	IRAK2	6
RARS	5	CBX3	5	KCTD18	6
RASSF8	5	CDC25A	5	LOC101927253	6
RFC2	5	CDC7	5	LOC102724156	6
RFC4	5	CDK2	5	LPCAT1	6
RNASEH2A	5	CDK4	5	MFSD11	6
RPS27L	5	CDKN3	5	MRPL9	6
SEC22B	5	CHAF1B	5	MTERF3	6
SERPINF1	5	CHMP7	5	NAT10	6
SET	5	CLECL1	5	PAPD4	6
SHCBP1	5	CMSS1	5	PAXIP1	6
SLF1	5	CTSE	5	PELO	6
SNRPA	5	DDX23	5	PFKFB3	6
SNRPD3	5	DIDO1	5	PHYKPL	6
SNRPG	5	DMD	5	PIK3CG	6
SPAG5	5	DMWD	5	PLCD1	6
SPOCK1	5	DNA2	5	SLC4A4	6
SSRP1	5	DONSON	5	SLC7A5	6
SSTR5	5	DPM1	5	SLCO4A1	6
STON1	5	DPYSL2	5	SND1	6
TACC3	5	DPYSL3	5	TIMP1	6
THBS2	5	EEF1E1	5	VARS	6
TIMM21	5	EIF3I	5	ZNF367	6
TLR3	5	ELAC2	5	ZNF593	6

GSE14333-USA		GSE14333-Melbourn		GSE4183-Normal colon	
Gene Name	Commonalities for all pathway members	Gene Name	Commonalities for all pathway members	Gene Name	Commonalities for all pathway members
TOR3A	5	EMC8	5	AATF	5
TSPYL5	5	ERICH1	5	ACBD6	5
TTI1	5	EZR	5	ACOT9	5
TUBB6	5	FAM217B	5	ACTN1	5
UBE2T	5	FANCD2	5	ADK	5
WDR7	5	FANCE	5	ANKRD12	5
WIBG	5	FARSA	5	ATAD2	5
ZCCHC2	5	FAS	5	BCL2L12	5
ZDHHC9	5	FBLN1	5	C2CD4A	5
ZEB2	5	FBLN5	5	C2orf88	5
ZFPM2	5	FBN1	5	C9orf16	5
ZMAT3	5	FSTL1	5	CBFB	5
ABAT	4	GAR1	5	CD276	5
ABHD3	4	GJB3	5	CD44	5
ACAT2	4	GNB4	5	CDCA2	5
ADAMTS1	4	GRPEL1	5	CEBPB	5
ADPRHL2	4	GSE1	5	CES2	5
ADRA1D	4	HAT1	5	CHTF18	5
AEN	4	HEG1	5	CNPY2	5
AFG3L2	4	HLX	5	CRNDE	5
ANGPTL2	4	HMMR	5	CXCL1	5
ANKFY1	4	IFT52	5	CXCL16	5
APOL1	4	IL1R1	5	DDX50	5
ARFGAP1	4	INTS10	5	DHX30	5
ASCL2	4	ISG20L2	5	DONSON	5
ASF1B	4	ITGA3	5	E2F3	5

GSE14333-USA		GSE14333-Melbourn		GSE4183-Normal colon	
Gene Name	Commonalities for all pathway members	Gene Name	Commonalities for all pathway members	Gene Name	Commonalities for all pathway members
BASP1	4	KCNF1	5	EIF3B	5
BCAT1	4	KIAA1462	5	ENC1	5
BGN	4	KNSTRN	5	ENTPD5	5
BICC1	4	KRT8	5	EXOSC8	5
BID	4	LAMA3	5	FGD6	5
BOC	4	LMNA	5	FIBP	5
BVES	4	LMO7	5	GEMIN5	5
C11orf95	4	LOC101929450	5	GRINA	5
C18orf8	4	LOC102724434	5	HEATR1	5
C1orf112	4	LOC105369167	5	HILPDA	5
C1QBP	4	LOC105372881	5	HNRNPD	5
C1R	4	LOC150005	5	LOC101929340	5
C22orf31	4	LOC221122	5	MAD2L2	5
C5AR1	4	MAP2K4	5	MAP4K4	5
CACYBP	4	MED17	5	METTL7A	5
CALD1	4	MIS18A	5	MSANTD3	5
CCDC185	4	MORC4	5	MTFR2	5
CCDC27	4	MPDU1	5	NANP	5
CCDC34	4	MPDZ	5	NINJ1	5
CCDC68	4	MRPL19	5	NOC3L	5
CCNE2	4	MRPL3	5	NOC4L	5
CCNF	4	MRPS16	5	OSBPL3	5
CCNL1	4	MSRA	5	PFDN2	5
CDC123	4	NDUFV2	5	PFKM	5
CDC25C	4	NFE2L3	5	PLP1	5
CDK14	4	NMT1	5	POLA1	5

GSE14333-USA		GSE14333-Melbourn		GSE4183-Normal colon	
Gene Name	Commonalities for all pathway members	Gene Name	Commonalities for all pathway members	Gene Name	Commonalities for all pathway members
CENPF	4	NOLC1	5	POLB	5
CENPU	4	NOP58	5	POLD1	5
CEP76	4	NOVA2	5	POLR1C	5
CLIC4	4	NUP50	5	POP7	5
CLTB	4	NXF1	5	PPP2R3A	5
CNGB1	4	OR5L2	5	PRR7	5
COL1A2	4	OSMR	5	PSRC1	5
COL3A1	4	PDE1A	5	R3HDM1	5
COL6A1	4	PDIA4	5	RBBP7	5
COL6A2	4	PES1	5	RCN1	5
COL6A3	4	PLA2G4A	5	RPP40	5
COL8A1	4	PLAGL2	5	RRN3	5
CPNE1	4	PLEKHN1	5	RRP1B	5
CTGF	4	PMAIP1	5	SAP30	5
CXXC1	4	POLA2	5	SCARA5	5
CYR61	4	POP7	5	SF3B3	5
CYSTM1	4	PPIF	5	SH3PXD2B	5
DBI	4	PPP1CA	5	SHISA5	5
DDX21	4	PSMC6	5	SKA3	5
DDX54	4	PTRF	5	SLC3A2	5
DENND5A	4	PTTG1	5	SLC6A6	5
DLC1	4	QPRT	5	SMYD2	5
DNAJC5	4	RAB11A	5	SNRPC	5
DNMT1	4	RAD54L	5	SUPT16H	5
DOCK5	4	RAE1	5	TALDO1	5
DPM1	4	RBM15	5	TBC1D16	5

GSE14333-USA		GSE14333-Melbourn		GSE4183-Normal colon	
Gene Name	Commonalities for all pathway members	Gene Name	Commonalities for all pathway members	Gene Name	Commonalities for all pathway members
DSCC1	4	RPA1	5	TM2D2	5
DSE	4	RRP9	5	TMEM185B	5
DSG2	4	RUNX1T1	5	TMEM206	5
DTYMK	4	RUVBL1	5	TNS4	5
DUSP4	4	SARNP	5	TRIM28	5
DZIP1	4	SDHB	5	VASN	5
E2F3	4	SEMA5A	5	WDR3	5
ECM2	4	SERPINF1	5	XPOT	5
EFEMP2	4	SERPING1	5	ZMYND19	5
EIF4G1	4	SFN	5	ZNF280C	5
EIF6	4	SFRP2	5	ZNF623	5
ERCC6L	4	SH3BGRL3	5	ZPR1	5
ERI2	4	SLAMF6	5	ADAM10	4
EZH2	4	SLBP	5	AGTRAP	4
F12	4	SLC39A6	5	ARL6IP6	4
FANCD2	4	SLC6A14	5	ARV1	4
FBN1	4	SMARCA4	5	ASPHD1	4
FDPS	4	SMC4	5	ATF1	4
FLNA	4	SPAG7	5	ATG4A	4
FLVCR1	4	SPAG8	5	ATL2	4
FSTL1	4	SPARCL1	5	BRIX1	4
GAS2	4	SPC25	5	BUB3	4
GBP2	4	SPCS3	5	C4BPB	4
GCN1	4	STIL	5	C4orf46	4
GCSH	4	STX8	5	CAD	4
GFPT2	4	SVEP1	5	CARNMT1	4

GSE14333-USA		GSE14333-Melbourn		GSE4183-Normal colon	
Gene Name	Commonalities for all pathway members	Gene Name	Commonalities for all pathway members	Gene Name	Commonalities for all pathway members
GGT5	4	SYT7	5	CCDC138	4
GJB3	4	TAF4	5	CCDC86	4
GLIS2	4	TAGLN	5	CDC25B	4
GMPS	4	TCFL5	5	CDKN2B	4
GNB4	4	THOC6	5	CENPW	4
GPC6	4	TMEM151B	5	CETN2	4
GREM2	4	TMEM160	5	CHGA	4
GTF2A2	4	TNFRSF10B	5	CHP2	4
GTF2IRD1	4	TNS1	5	CHSY1	4
GTPBP4	4	TOP2A	5	CIRH1A	4
H2AFX	4	TPX2	5	CLK1	4
HAT1	4	TSPAN6	5	CMTR2	4
HIST2H2AA3	4	TTF2	5	CNTN3	4
HK2	4	TTLL12	5	COG6	4
HMMR	4	TUBB6	5	COPS8	4
HP	4	TUBG1	5	CXCL2	4
HPSE	4	TVP23B	5	CXCL3	4
HSPA9	4	UBA6	5	DCLRE1A	4
HSPH1	4	UNG	5	DDIAS	4
HTRA1	4	VPS37A	5	DIAPH3	4
HUS1B	4	VPS37B	5	DIEXF	4
IER5	4	YTHDF1	5	DLG5	4
IGFN1	4	ZMAT3	5	DNAJC21	4
INHBA	4	ZNF121	5	DOCK10	4
INTS2	4	ZNF843	5	DYSF	4
IWS1	4	ACACB	4	EDN3	4

GSE14333-USA		GSE14333-Melbourn		GSE4183-Normal colon	
Gene Name	Commonalities for all pathway members	Gene Name	Commonalities for all pathway members	Gene Name	Commonalities for all pathway members
JAK2	4	ACADVL	4	EFNA4	4
KCNE4	4	ACSL6	4	ELF1	4
KCNH5	4	ADSL	4	ENAH	4
KCNT1	4	AEBP1	4	EPHX2	4
KCTD5	4	AFAP1-AS1	4	EPHX4	4
KCTD9	4	AGMAT	4	ETV4	4
KDSR	4	AGPAT5	4	EXOSC5	4
KIAA1524	4	AHR	4	FAAP20	4
KIAA1644	4	AK2	4	FAM83D	4
KIF15	4	AKT3	4	FANCG	4
KIF1C	4	AMD1	4	FARSA	4
KIF20A	4	AMPH	4	FBRSL1	4
KIF22	4	ANKRD40	4	FBXO5	4
KIFC1	4	ANKRD49	4	FGD5	4
KLHDC1	4	ANO1	4	FGFR2	4
KLHL5	4	ANTXR2	4	G6PD	4
KLK13	4	ANXA1	4	GART	4
KNTC1	4	ANXA10	4	GCNT3	4
LAMA4	4	AOC3	4	GDF15	4
LAYN	4	APEH	4	GEMIN2	4
LDB2	4	ARHGEF6	4	GGTA1P	4
LGALS3BP	4	ARL4C	4	GIMAP4	4
LLPH	4	ASAH1	4	GLG1	4
LMNB1	4	ASUN	4	GMCL1	4
LOC100129476	4	ATAD1	4	GOLGA2P5	4
LOC100506142	4	ATG101	4	GRIN2D	4

GSE14333-USA		**GSE14333-Melbourn**		**GSE4183-Normal colon**	
Gene Name	Commonalities for all pathway members	Gene Name	Commonalities for all pathway members	Gene Name	Commonalities for all pathway members
LOC101927040	4	ATP8B2	4	HAUS6	4
LOC101929144	4	ATP9A	4	HDAC2	4
LOC340107	4	AXIN2	4	HNRNPH3	4
LOX	4	AXL	4	HPGDS	4
LOXL1	4	BCL10	4	IER3	4
MBD2	4	BCL6	4	IFITM2	4
MCMBP	4	BHMT2	4	IFITM3	4
ME2	4	BNC2	4	IL1RN	4
MED20	4	BTNL9	4	IL6	4
MEIS1	4	C10orf10	4	INTS5	4
MLLT11	4	C10orf2	4	IPO5	4
MMP24-AS1	4	C19orf33	4	JAG2	4
MORN1	4	C19orf84	4	KCMF1	4
MRC2	4	C3orf52	4	KCTD20	4
MRPL14	4	CALCOCO1	4	KDM1A	4
MRPL35	4	CALD1	4	KIF15	4
MRPL37	4	CAMTA2	4	KIFC1	4
MRPL46	4	CAPG	4	KLK11	4
MRPS27	4	CCAR2	4	LAS1L	4
MXD1	4	CCDC80	4	LGR5	4
MXRA5	4	CCDC88A	4	LIN7C	4
NCAPG	4	CCNE2	4	LOC440792	4
NCAPG2	4	CD55	4	LPAR6	4
NDN	4	CDC27	4	LPGAT1	4
NEK4	4	CDC42EP1	4	MAP4K3	4
NFE2L1	4	CDCA4	4	MED27	4

GSE14333-USA		**GSE14333-Melbourn**		**GSE4183-Normal colon**	
Gene Name	Commonalities for all pathway members	Gene Name	Commonalities for all pathway members	Gene Name	Commonalities for all pathway members
NNMT	4	CDH22	4	MLST8	4
NOL4L	4	CDK14	4	MORC4	4
NOTCH3	4	CEBPZ	4	MPEG1	4
NRBF2	4	CENPN	4	MYC	4
NTM	4	CEP78	4	NBEAL2	4
NUAK1	4	CFAP20	4	NHP2	4
NUP107	4	CLEC1A	4	NLE1	4
NUP205	4	CNOT7	4	NUDT1	4
NUP37	4	COL14A1	4	NXT1	4
NUTF2	4	COL6A2	4	ORC2	4
OR10H1	4	COX7A1	4	OSTM1	4
OSER1	4	CPSF3L	4	OTUB2	4
PAK6	4	CRIP1	4	OXR1	4
PCDH7	4	CRISPLD2	4	PADI2	4
PDGFC	4	CRYAB	4	PAK1IP1	4
PDLIM2	4	CRYZ	4	PCDH19	4
PDLIM3	4	CSGALNACT2	4	PDCD2L	4
PHLDA1	4	CTGF	4	PFAS	4
PHLDA2	4	CX3CR1	4	PHC2	4
PIGW	4	CXCL16	4	PHF3	4
PKD2	4	CYP1A1	4	PLXNA1	4
PKP2	4	CYR61	4	POLR1E	4
PLCG1	4	CYSTM1	4	PPIH	4
PLCL2	4	D21S2088E	4	PPP1CC	4
PLXNB2	4	DCLK1	4	PPP4R3B	4
PNPT1	4	DDR2	4	PRDX1	4

GSE14333-USA		GSE14333-Melbourn		GSE4183-Normal colon	
Gene Name	Commonalities for all pathway members	Gene Name	Commonalities for all pathway members	Gene Name	Commonalities for all pathway members
POGK	4	DIP2B	4	PRIMPOL	4
POLDIP2	4	DYM	4	PTAFR	4
POLQ	4	ECM2	4	PTCD3	4
POLR2E	4	EFEMP2	4	PTPN22	4
PPDPF	4	EIF3D	4	PTRH2	4
PPIF	4	EPM2AIP1	4	PUS1	4
PPP1CA	4	EPOR	4	PUS7	4
PPP1R18	4	ESCO2	4	PXMP2	4
PPP1R3D	4	ETV5	4	RAD23A	4
PRLR	4	FAM129B	4	RBL2	4
PRPF31	4	FAM160B2	4	RCC1	4
PSMA1	4	FANCA	4	RILPL2	4
PSMA4	4	FANCF	4	RNF38	4
PSMB3	4	FBXL7	4	RPL22L1	4
PSMD12	4	FEM1A	4	RRP9	4
PSMD13	4	FEZ1	4	RSL1D1	4
PTGFRN	4	FGF13-AS1	4	SAPCD2	4
QPRT	4	FGFR1	4	SATB2	4
RAE1	4	FH	4	SEMA6A	4
RAI14	4	FHL1	4	SENP6	4
RANGAP1	4	FKBPL	4	SHB	4
RBM12	4	FLI1	4	SHPRH	4
RCN3	4	FLVCR1	4	SLC29A1	4
RGS2	4	FMO2	4	SLC9A9	4
RGS4	4	FOXD1	4	SMOX	4
RHOF	4	GAS1	4	SNTB1	4

GSE14333-USA		GSE14333-Melbourn		GSE4183-Normal colon	
Gene Name	Commonalities for all pathway members	Gene Name	Commonalities for all pathway members	Gene Name	Commonalities for all pathway members
RHOQ	4	GBX2	4	SOCS3	4
RNF34	4	GEM	4	SPOPL	4
RPRD1B	4	GFRA1	4	SRP9	4
RTN4IP1	4	GLOD4	4	SSX2IP	4
S100A11	4	GNAO1	4	STMN1	4
SAC3D1	4	GPR22	4	STX12	4
SARNP	4	GPR26	4	TAF1A	4
SART3	4	GSR	4	TEAD4	4
SCARB1	4	GTF2E2	4	TESC	4
SDC2	4	GTF2F1	4	TEX10	4
SDF2L1	4	GYPC	4	TLK1	4
SERINC3	4	HDAC1	4	TM7SF3	4
SERPINH1	4	HNRNPK	4	TMEM147	4
SGCB	4	HSPA9	4	TMEM72	4
SLC22A6	4	HSPB7	4	TPD52L2	4
SLC24A4	4	HSPE1	4	TRAF3IP3	4
SLC35C2	4	HTR1E	4	TRAPPC13	4
SLC39A6	4	IBTK	4	TRBC1	4
SLC9A1	4	ICT1	4	TRMT13	4
SLIT3	4	IFRD2	4	TSPAN7	4

Complete Tables for Comparative Analysis in Chapter 5

TCGA-COADREAD Project

Concor dant predict ors for both cohorts	Frequency	P-Value	Unique Predictors		Unique predictors			
ADNP	5	2.007E-11	TCGA-Colorectal-WTTP53	Frequency	Pvalue	TCGA-ColorectalMutantTP53	Frequency	Pvalue
AURKA	8	1.46E-06	BOLA3	6	0.0450525	C5orf41	9	0.0035863
AURKB	6	0.0024815	CBFA2T2	5	7.825E-08	CHAF1B	8	0.0339956
BUB1B	6	0.0151441	CHD6	5	8.831E-08	ERCC6L	8	0.0040474
CCDC86	5	0.0035881	DZIP1	5	0.0032789	KIAA0101	8	0.0014184
CCNA2	8	0.0099209	POP5	5	0.0005303	SERINC1	7	0.0182255
CCNB2	6	0.0366	RPL22L1	5	5.319E-20	TRAIP	7	0.0337

Concor dant predict ors for both cohorts	Frequency	P- Value	Unique Predictors		Unique predictors			
		538						73
CCT2	7	0.0068 801	AP1S3	5	0.0020607	AURKAIP1	6	0.0007 656
CDC20	5	0.0023 573	ARHGEF11	6	0.0321763	BMPR2	6	0.0023 833
CDCA5	8	0.0402 606	CDCA2	5	1.188E-07	CLCN7	6	2.075E -14
CDCA8	5	0.0352 498	EIF2S1	5	0.0011267	KIF20A	6	0.0447 28
CDKN3	5	0.0361 064	GEMIN7	6	0.0059802	MYBL2	6	1.089E -08
DBF4	5	0.0294 442	GPX1	5	0.0220624	NAP1L3	6	0.0384 559
DSCC1	7	0.0009	KIAA2026	5	0.0015767	PGAM5	6	0.0264

Concor dant predict ors for both cohorts	**Frequency**	**P-Value**	**Unique Predictors**		**Unique predictors**			
		823						906
E2F1	8	1.039E-07	LOC647979	5	1.797E-14	PGR	6	0.0180103
FAM54A	7	0.0426078	NR3C2	6	0.0020093	PSAT1	6	0.0350036
FANCB	6	0.0093759	PBK	7	2.044E-10	RAB27A	6	2.732E-06
H2AFZ	5	0.000219	PCDH19	5	0.0005507	RAD51	6	0.0172153
KIF4A	6	0.0023102	PHF20	5	9.626E-13	RIF1	6	0.0451558
MND1	7	0.0366204	POFUT1	5	3.466E-11	SECISBP2L	6	0.0013716
NUP37	7	0.0212674	RNF19B	5	3.396E-07	TELO2	6	0.0256418

Concor dant predict ors for both cohorts	Frequency	P- Value	Unique Predictors		Unique predictors			
ORC1L	6	0.0320302	SKA1	6	0.0001044	ZBTB4	6	0.0140555
ORC6L	8	0.0029739	AEN	5	4.063E-11	ACD	5	2.142E-05
PA2G4	7	0.0294297	AKAP13	5	0.0130803	ATP5F1	5	2.692E-05
PNPT1	8	0.0124803	ASH1L	5	0.0088427	BOC	5	0.0033615
RAN	6	0.0062796	ATP5A1	5	1.143E-18	BRCA1	5	0.0089958
RCC1	5	0.0281494	ATRX	5	0.0010952	C13orf33	5	0.0190404
RFC4	7	0.0316941	BAT2	6	0.0153324	C17orf86	5	7.316E-06
RFC5	8	0.0190186	C15orf23	7	0.0177738	CBX7	5	0.0139492

Concor dant predict ors for both cohorts	Frequency	P-Value	Unique Predictors		Unique predictors			
RRM2	7	0.0075561	C3orf26	7	0.0142335	CCDC150	5	0.0025827
SKA3	5	9.956E-07	C4orf46	6	0.0372259	CDK10	5	0.0050387
SNRPF	6	0.03248	C6orf120	5	0.0183539	CENPI	5	0.0415519
SPC25	7	0.0422415	CCDC8	5	0.0326559	CHPF	5	0.0062187
TPX2	6	3.767E-07	CDKN1A	5	3.352E-13	DUSP4	5	7.09E-13
TTK	6	0.0210373	CEBPB	5	2.696E-08	ECT2	5	0.0428281
UBE2C	9	2.423E-08	CENPM	5	0.0019845	EDEM3	5	0.0001448
XRCC2	6	0.0003146	COPE	5	0.0065805	EVC2	5	0.0387312

Concor dant predict ors for both cohorts	Frequency	P-Value	Unique Predictors		Unique predictors			
			DDAH2	6	3.618E-06	FCHO2	5	0.0009199
			DLGAP5	5	0.0020723	GFI1	5	1.474E-07
			DPY19L4	5	2.186E-07	GOLPH3L	5	7.392E-06
			EFNA4	5	0.0235435	KIF11	5	0.0337183
			EFS	5	0.0240362	KIF15	5	0.0109783
			ERH	6	0.0024016	LENG8	5	0.0013141
			FASTKD1	5	0.0243502	LIMA1	5	3.326E-09
			FBXO5	5	0.0026209	MAP3K3	5	0.0079

Concor dant predict ors for both cohorts	Frequency	P-Value	Unique Predictors		Unique predictors			
								33
			FDXR	8	4.772E-18	MTHFD2	5	0.0020327
			HECA	5	2.002E-05	NCAPD3	5	0.0244773
			HNRNPC	5	7.563E-07	NCAPG2	5	0.0015963
			KIAA0182	5	0.0280741	OGFR	5	3.816E-05
			KIAA0240	5	7.369E-05	PEG3	5	0.01992
			MCM7	5	0.0170605	PHLDA1	5	0.0006578
			MDM2	6	6.77E-31	PIK3CA	5	0.0003526
			MLL3	5	0.0340408	POC1A	5	0.0364

Concor dant predict ors for both cohorts	Frequency	P-Value	Unique Predictors		Unique predictors			
								815
			MPDU1	5	1.987E-21	PRC1	5	0.0421 074
			MRPL35	5	5.25E-05	PRDM6	5	0.0083 925
			NIPBL	5	0.0117857	PRICKLE1	5	0.0072 093
			PRMT1	6	0.0006633	SETD1A	5	5.104E -05
			PSMA3	6	0.01839	SGOL2	5	0.0194 816
			PSMA6	5	0.0005269	SHMT2	5	6.085E -05
			PSME2	5	9.399E-05	TACC3	5	0.0149 187
			RPAP3	5	0.0267138	TOP2A	5	0.0133

Concor dant predict ors for both cohorts	**Frequency**	**P- Value**	**Unique Predictors**		**Unique predictors**			
								41
			RPS27L	7	5.028E-27	TXLNA	5	0.0042 055
			RUVBL2	7	0.0007353	XPOT	5	0.0038 467
			SLC25A22	5	0.0198328	ZEB1	5	0.0019 442
			SNRPD1	5	5.039E-05	ZFYVE1	5	0.0346 335
			SNX30	6	0.0161471	ZGPAT	5	1.721E -13
			SP4	5	0.0191079	MDM2		6.77E- 31
			SPATA18	5	8.302E-43	DDB2		1.016E -25
			TIPIN	5	0.0013685	FAS		8.31E-

Concor dant predict ors for both cohorts	Frequency	P-Value	Unique Predictors		Unique predictors			
								22
			TNFRSF10B	7	3.114E-10	CDKN1A		3.352E-13
			TNFSF9	7	4.251E-11	ZMAT3		1.821E-12
			TOMM22	5	0.0065802	SIAH1		2.3E-10
			TRIAP1	7	2.486E-10	TNFRSF10B		3.114E-10
			UBE2N	7	0.0032433	BAX		3.306E-10
			UBL3	5	2.863E-08	CCNG1		2.395E-07
			UQCRFS1	5	1.558E-06	RPRM		4.788E-07
			UQCRQ	5	0.0139214	CASP3		7.635E

Concor dant predict ors for both cohorts	Frequency	P-Value	Unique Predictors		Unique predictors			
								-07
			WARS	5	0.0058847	ATR		8.533E-07
			ZNF445	5	0.0040042	BCL2L1		1.925E-06
			ZZEF1	6	0.0002679	GADD45A		2.917E-06
			DDB2		6.77E-31	TP53		1.496E-05
			FAS		1.016E-25	PPM1D		4.132E-05
			ZMAT3		8.31E-22	BBC3		7.902E-05
			SIAH1		3.352E-13	BCL2		0.000218
			BAX		1.821E-12	ATM		0.0003

Concor dant predict ors for both cohorts	Frequency	P- Value	Unique Predictors		Unique predictors			
								136
			CCNG1		2.3E-10	TP53AIP1		0.0003 465
			RPRM		3.114E-10	CDKN2A		0.0006 674
			CASP3		3.306E-10	PIGS		0.0009 318
			ATR		2.395E-07	STEAP3		0.0013 274
			BCL2L1		4.788E-07	TSC2		0.0026 156
			GADD45A		7.635E-07	SESN1		0.0032 972
			TP53		8.533E-07	SFN		0.0097 301
			PPM1D		1.925E-06	APAF1		0.0111

Concor dant predict ors for both cohorts	Frequency	P-Value	Unique Predictors		Unique predictors			
								702
			BBC3		2.917E-06	GADD45B		0.0130177
			BCL2		1.496E-05	GADD45G		0.0220574
			ATM		4.132E-05	RRM2B		0.0435826
			TP53AIP1		7.902E-05	PERP		0.0480771
			CDKN2A		0.000218			
			PIGS		0.0003136			
			STEAP3		0.0003465			
			TSC2		0.0006674			
			SESN1		0.0009318			
			SFN		0.0013274			

Concor dant predict ors for both cohorts	Frequency	P- Value	Unique Predictors		Unique predictors			
			APAF1		0.0026156			
			GADD45B		0.0032972			
			GADD45G		0.0097301			
			RRM2B		0.0111702			
			PERP		0.0130177			
					0.0220574			
					0.0435826			
					0.0480771			

TCGA-STAD Project

Concordant predictors for Both cohorts		Unique predictors for the TCGA-STAD- MutantTP53			Unique predictors for the TCGA-STAD-Wild TypeTP53		
Gene Name	P-value	Gene Name	Frequency	P-value	Gene Name	Frequency	P-value
BUB1	1.3425E-51	EXO1	9	3.5567E-06	SNORD115-17	19	4.6176E-39
CDCA8	7.7448E-50	CCNA2	8	3.0602E-12	BARX1	17	5.4463E-16
CDCA3	4.1505E-41	CENPA	8	3.9839E-09	MUC13	16	2.7321E-47
POLE2	1.0669E-49	DSCC1	8	3.6864E-33	TUBG2	16	9.6025E-43
CENPF	1.818E-58	DTL	8	1.2968E-17	ADCY8	15	6.0706E-41
FOXM1	5.1384E-43	ERCC6L	8	1.074E-43	FERMT1	14	3.7921E-38
FAM54A	1.5191E-49	KIF18B	8	5.2841E-06	SNORD115-41	13	3.1906E-46
CDCA2	6.0297E-33	MAD2L1	8	1.5134E-10	FLJ42393	13	1.3841E-45

Concordant predictors for Both cohorts		Unique predictors for the TCGA- STAD- MutantTP53			Unique predictors for the TCGA- STAD- Wild TypeTP53		
Gene Name	P-value	Gene Name	Frequency	P-value	Gene Name	Frequency	P-value
CASC5	4.1383E-50	ORC1L	8	2.003E-36	CNPY2	13	7.1919E-42
C12orf48	6.8986E-45	PRIM1	8	4.4905E-13	GRINA	13	1.035E-40
BUB1B	1.3428E-50	RFC3	8	0.00418791	TRIM15	12	8.3554E-47
NCAPH	3.4128E-46	RNF150	8	0.01926792	TCEA2	12	4.5471E-45
RRM2	4.1269E-35	SPAG5	8	3.2241E-15	SNORD29	12	1.8176E-33
UBE2C	4.0639E-62	TRIP13	8	2.3199E-10	LPP	12	1.5986E-32
CCNF	3.7804E-23	TROAP	8	6.0609E-05	SNORA36C	12	2.1564E-13
NCAPG	4.0629E-48	AHCTF1	7	7.4581E-17	ZNF559	11	4.8751E-46
CDC25C	2.2008E-29	ATP8B2	7	1.3856E-15	ESRP1	11	1.0525E-45
SKA1	2.3217E-34	AURKA	7	0.00026095	FA2H	11	2.0893E-45

Concordant predictors for Both cohorts		**Unique predictors for the TCGA- STAD- MutantTP53**			**Unique predictors for the TCGA- STAD- Wild TypeTP53**		
Gene Name	P-value	Gene Name	Frequency	P-value	Gene Name	Frequency	P-value
CDC20	1.1016E-40	C10orf72	7	0.00046173	TRIM31	11	4.3576E-43
CDCA5	2.4148E-31	C11orf82	7	1.5658E-06	VEGFB	11	1.0008E-39
FBN1	2.8004E-11	C1orf112	7	5.4377E-06	GPR35	11	1.5679E-39
GSG2	6.5295E-41	C21orf45	7	1.9437E-19	AGR2	11	1.0475E-30
NEK2	1.6142E-36	CCNB2	7	1.7886E-15	FBLL1	11	1.967E-30
PRR11	1.8375E-44	CDC25A	7	7.3221E-30	BNIPL	11	4.7807E-11
CENPO	5.2289E-44	CDC45	7	2.743E-23	CLSPN	10	6.4374E-47
NUF2	1.5191E-49	CDC6	7	2.4441E-12	C1orf106	10	3.9807E-46
CCNB1	6.0297E-33	CENPL	7	3.0756E-05	TMC5	10	2.3672E-45
HJURP	4.1383E-50	EPHA2	7	4.0178E-09	DNAJB12	10	6.0657E-44

Concordant predictors for Both cohorts		Unique predictors for the TCGA- STAD- MutantTP53			Unique predictors for the TCGA- STAD- Wild TypeTP53		
Gene Name	P-value	Gene Name	Frequency	P-value	Gene Name	Frequency	P-value
KIF23	6.8986E-45	FXYD6	7	2.401E-13	OPA3	10	5.0223E-43
KIFC1	1.3428E-50	KIF18A	7	5.4225E-09	OPA1	10	1.9116E-40
POLQ	3.4128E-46	KIF2C	7	6.5373E-05	ZNF720	10	5.5681E-37
PRC1	4.1269E-35	MELK	7	2.3731E-19	GPR78	10	9.6146E-37
RAD51AP1	4.0639E-62	MPDZ	7	1.3125E-29	CDCA7	10	5.2396E-32
RAD54L	3.7804E-23	OIP5	7	7.5097E-12	CXCL3	10	2.1508E-28
CENPE	4.0629E-48	ORC6L	7	1.3327E-28	CELF4	10	1.3186E-27
CHEK1	1.3425E-51	PLK4	7	5.2521E-29	BCL2L15	10	1.7917E-27
DNA2	7.7448E-50	RACGAP1	7	2.7876E-49	ZBTB20	10	1.6843E-26
EVPL	4.1505E-41	RAD51	7	1.3041E-16	SLC6A13	10	4.2531E-26
HMMR	1.0669E-49	RUNX1T1	7	3.7998E-23	CCL11	10	1.3525E-17
SASS6	1.818E-58	SGOL1	7	2.3999E-19	KRT31	10	1.4111E-10

Concordant predictors for Both cohorts		Unique predictors for the TCGA- STAD- MutantTP53			Unique predictors for the TCGA- STAD- Wild TypeTP53		
Gene Name	P-value	Gene Name	Frequency	P-value	Gene Name	Frequency	P-value
UBE2T	5.1384E-43	SKA3	7	3.9173E-12	DMRTC1	10	1.2485E-24
ASPM	6.0297E-33	SPC25	7	9.2596E-19	C9orf119	10	1.8371E-05
CKAP2L	4.1383E-50	STON1	7	3.1914E-11	POU2F1	9	3.5411E-47
FBXO5	6.8986E-45	TPX2	7	1.816E-13	C6orf222	9	1.5265E-45
KIF15	1.3428E-50	TTK	7	6.076E-07	POF1B	9	5.3732E-40
SERPINB5	3.4128E-46	WDR67	7	5.9215E-23	ZNF2	9	1.5978E-39
WDHD1	4.1269E-35	ZNF644	7	3.2632E-08	PITX1	9	7.8332E-39
		ANXA2	6	1.0235E-36	BCL11B	9	1.0066E-38
		BCL7C	6	1.0059E-45	TAOK2	9	2.3023E-38
		BDP1	6	1.5035E-31	GATA6	9	2.446E-38
		BLM	6	0.00054978	LOC84856	9	2.6701E-38
		BTAF1	6	2.9414E-09	TSPAN8	9	2.2564E-37
		C15orf42	6	0.00018525	LRRC48	9	1.8087E-34

Concordant predictors for Both cohorts		**Unique predictors for the TCGA- STAD- MutantTP53**			**Unique predictors for the TCGA- STAD- Wild TypeTP53**		
Gene Name	P-value	Gene Name	Frequency	P-value	Gene Name	Frequency	P-value
		C1orf135	6	0.00860724	MDH1B	9	4.4628E-29
		CASP8AP2	6	5.9829E-11	SFTPD	9	3.0273E-25
		CCDC138	6	5.1128E-12	SNHG11	9	4.1411E-19
		CCDC150	6	0.04621805	HIC1	9	9.4137E-19
		CCNT2	6	0.03360658	PPOX	9	0.00148945
		CEP97	6	4.3642E-05	ASPN	9	2.3248E-05
		CHAF1B	6	0.00097938	RADIL	9	2.3322E-23
		CKS2	6	0.04062223	GPR171	9	3.2166E-23
		CSE1L	6	0.0029036	C12orf27	8	3.2523E-47
		DCLK2	6	5.4898E-37	ELF3	8	2.6513E-46
		DHX9	6	0.00146327	FLJ44635	8	3.7368E-46
		DPYSL3	6	1.6065E-21	PRR15L	8	8.625E-46
		DSG3	6	9.4541E-13	MNX1	8	1.5628E-45

Concordant predictors for Both cohorts		Unique predictors for the TCGA- STAD- MutantTP53			Unique predictors for the TCGA- STAD- Wild TypeTP53		
Gene Name	P-value	Gene Name	Frequency	P-value	Gene Name	Frequency	P-value
		E2F1	6	3.4929E-14	PRSS3	8	2.1824E-45
		EZH2	6	4.597E-11	NFU1	8	5.9333E-45
		FAM108C1	6	7.9129E-15	HIST1H4C	8	1.6259E-43
		FAM72B	6	4.778E-07	GOLPH3L	8	3.232E-42
		FAM83H	6	1.4601E-11	FOXQ1	8	4.2607E-42
		FANCI	6	3.7976E-12	KDM4DL	8	7.7557E-42
		FBXL7	6	1.987E-09	GNPTG	8	1.4829E-40
		FEN1	6	1.0691E-16	CEP55	8	1.6875E-40
		FERMT2	6	1.0757E-09	DEPDC1B	8	3.4576E-38
		FOXN2	6	0.00052984	SF3B14	8	9.6828E-37
		GGT5	6	0.00015326	CENPV	8	1.9124E-34

Concordant predictors for Both cohorts		**Unique predictors for the TCGA-STAD- MutantTP53**			**Unique predictors for the TCGA-STAD-Wild TypeTP53**		
Gene Name	P-value	Gene Name	Frequency	P-value	Gene Name	Frequency	P-value
		KANK2	6	1.2938E-14	DSG2	8	2.088E-33
		KIAA1731	6	0.00019819	CAPSL	8	1.4321E-32
		KIF11	6	3.4725E-07	SCGB1D2	8	6.4631E-31
		KIF20A	6	7.7591E-19	HELB	8	4.4826E-29
		KNTC1	6	1.5722E-13	ANXA3	8	2.4519E-28
		KPNA2	6	0.00186681	GPRC5A	8	2.5052E-27
		LAD1	6	3.4488E-33	C8orf80	8	3.0975E-25
		MAP3K12	6	3.2903E-13	CASP10	8	6.2941E-22
		MAPK10	6	0.0302039	GABRQ	8	1.0238E-20
		MATR3	6	3.0935E-14	DDX25	8	1.3253E-20

Concordant predictors for Both cohorts		Unique predictors for the TCGA-STAD- MutantTP53			Unique predictors for the TCGA-STAD-Wild TypeTP53		
Gene Name	P-value	Gene Name	Frequency	P-value	Gene Name	Frequency	P-value
		MCM6	6	8.201E-17	FKBP10	8	4.3482E-14
		MPHOSPH9	6	5.144E-06	CRYGS	8	2.2736E-12
		MSRB3	6	1.2572E-12	MAGED4	8	7.9111E-11
		MYSM1	6	0.00733228	PDZD4	8	1.3295E-06
		NPAT	6	2.3334E-10	RIC3	8	4.2531E-26
		NUSAP1	6	1.2112E-11	GABRG3	8	9.7292E-11
		PHLDA2	6	0.00177364	HNF4A	7	5.3201E-47
		PIKFYVE	6	1.349E-23	GOLGA1	7	2.076E-46
		PKP3	6	2.2175E-32	NRARP	7	3.3828E-46
		PRKD1	6	3.1931E-20	ONECUT2	7	4.4355E-46

Concordant predictors for Both cohorts		**Unique predictors for the TCGA-STAD- MutantTP53**			**Unique predictors for the TCGA-STAD-Wild TypeTP53**		
Gene Name	P-value	Gene Name	Frequency	P-value	Gene Name	Frequency	P-value
		PRR24	6	1.095E-17	SPINT1	7	6.3069E-45
		RALY	6	0.00140715	HMGCLL1	7	1.1656E-44
		RANBP2	6	3.229E-13	HMG20A	7	1.2345E-43
		RECK	6	3.0064E-15	SHROOM3	7	1.939E-42
		RFC5	6	5.2101E-42	CHST1	7	1.0848E-41
		RIF1	6	5.9681E-15	NKIRAS1	7	6.7645E-40
		SETBP1	6	1.1133E-10	ANKS4B	7	1.2544E-39
		SGOL2	6	1.5177E-06	SNORD127	7	3.9725E-39
		SLIT2	6	1.0943E-36	C11orf2	7	1.217E-37
		SYNE1	6	9.3895E-35	SEL1L3	7	1.4377E-37

Concordant predictors for Both cohorts		**Unique predictors for the TCGA-STAD- MutantTP53**			**Unique predictors for the TCGA-STAD-Wild TypeTP53**		
Gene Name	P-value	Gene Name	Frequency	P-value	Gene Name	Frequency	P-value
		TIMELESS	6	1.6328E-37	DNAJC3	7	1.9359E-36
		TNS1	6	8.0724E-17	FAM83B	7	2.844E-35
		TOP2A	6	6.9523E-08	SNX22	7	4.8823E-35
		TRAIP	6	5.893E-08	SEMA4G	7	5.8881E-35
		UHMK1	6	1.7527E-37	SNORD116-25	7	9.3654E-35
		UHRF1	6	3.0294E-25	PIP5K1B	7	1.1898E-34
		USP24	6	0.01024892	CAPN8	7	3.0244E-34
		XPO1	6	9.5331E-07	ZCCHC3	7	2.7735E-33
		ZC3H11A	6	7.866E-21	SLC22A17	7	7.9259E-33
		ZCCHC24	6	4.9497E-38	PKP2	7	2.487E-32

Concordant predictors for Both cohorts		**Unique predictors for the TCGA- STAD- MutantTP53**			**Unique predictors for the TCGA- STAD- Wild TypeTP53**		
Gene Name	P-value	Gene Name	Frequency	P-value	Gene Name	Frequency	P-value
		ZWINT	6	7.179E-30	C10orf119	7	7.5345E-31
		CCNE1	6	7.1464E-17	ADAMTS6	7	3.9704E-30
		CDK1	6	5.5099E-26	FAM128B	7	6.615E-30
		ABCC9	5	0.00432291	SLC41A2	7	1.1447E-29
		ADAMTSL3	5	6.4158E-23	TRAM1L1	7	9.5035E-28
		ADAT2	5	2.5421E-11	TRIM47	7	2.6082E-25
		AKT3	5	2.8481E-19	SRPK1	7	8.2622E-25
		ALS2CL	5	6.6737E-12	SMYD5	7	8.1796E-22
		ANGPTL1	5	0.0040671	DNAJB2	7	5.4576E-20
		ANK2	5	8.5208E-05	C2orf62	7	1.5875E-18

Concordant predictors for Both cohorts		Unique predictors for the TCGA-STAD- MutantTP53			Unique predictors for the TCGA-STAD-Wild TypeTP53		
Gene Name	P-value	Gene Name	Frequency	P-value	Gene Name	Frequency	P-value
		ANXA2P2	5	3.69E-19	REL	7	2.1304E-18
		ARHGAP20	5	2.9075E-09	SLC37A1	7	2.7649E-18
		ATAD2	5	2.3061E-19	ANKRD36BP1	7	1.2305E-16
		AXL	5	1.0266E-08	PEX1	7	3.2951E-11
		B3GALNT2	5	9.0829E-26	PTTG1	7	1.3739E-09
		BIRC5	5	2.5043E-32	FANCA	7	7.9661E-09
		BOC	5	9.0569E-19	XAGE1D	7	9.1804E-07
		BRCA1	5	0.00126478	COTL1	7	1.0145E-06
		BRIP1	5	2.0437E-07	FILIP1	7	1.1302E-21
		C17orf53	5	1.2002E-17	NTF4	7	7.4178E-41

Concordant predictors for Both cohorts		Unique predictors for the TCGA- STAD- MutantTP53			Unique predictors for the TCGA- STAD- Wild TypeTP53		
Gene Name	P-value	Gene Name	Frequency	P-value	Gene Name	Frequency	P-value
		C1R	5	1.1896E-05	SGCD	7	3.8894E-11
		C1S	5	3.9476E-10	GINS4	7	6.681E-06
		CALD1	5	1.0098E-19	SAMHD1	7	2.3318E-05
		CAND1	5	9.3314E-05	OR4C16	7	1.1716E-35
		CAPN1	5	1.2843E-06	IFFO1	6	7.3575E-47
		CCDC14	5	0.01638143	ANKRD9	6	8.9441E-47
		CDCP1	5	1.924E-05	EPS8L3	6	9.9547E-47
		CDKN3	5	1.3254E-25	RPS18	6	2.8607E-45
		CDS1	5	1.6298E-12	MAPK11	6	7.4859E-45
		CEP350	5	4.0984E-07	E2F8	6	7.5018E-45
		CLIP3	5	1.0928E-17	PLS1	6	8.8202E-45

Concordant predictors for Both cohorts		Unique predictors for the TCGA- STAD- MutantTP53			Unique predictors for the TCGA- STAD- Wild TypeTP53		
Gene Name	P-value	Gene Name	Frequency	P-value	Gene Name	Frequency	P-value
		COMMD4	5	0.03936823	UBXN6	6	1.7729E-44
		DENND4A	5	0.00057595	PFDN5	6	1.936E-43
		DEPDC1	5	2.498E-27	PRR15	6	4.8833E-43
		DLGAP5	5	1.1943E-30	MGC4473	6	5.2237E-43
		E2F7	5	1.0429E-20	GLTP	6	1.4716E-42
		EFEMP1	5	3.8094E-07	ARHGAP28	6	5.8891E-41
		ELMO3	5	3.8566E-40	HNF1A	6	1.2334E-40
		EME1	5	9.0187E-15	PKNOX2	6	5.3636E-40
		EPS8L2	5	0.0008603	PLA2G2C	6	2.2035E-39
		ESCO2	5	3.0771E-34	STK24	6	4.277E-39
		EVC	5	2.0042E-09	OVOL2	6	8.8004E-39
		EXOSC8	5	0.00256876	MRPL46	6	1.3429E-38
		FAM72A	5	1.2749E-13	RBM14	6	1.553E-38

Concordant predictors for Both cohorts		**Unique predictors for the TCGA- STAD- MutantTP53**			**Unique predictors for the TCGA- STAD- Wild TypeTP53**		
Gene Name	P-value	Gene Name	Frequency	P-value	Gene Name	Frequency	P-value
		FANCD2	5	0.01404021	KHDRBS3	6	4.1225E-38
		FBXO11	5	0.00175176	FUT2	6	7.743E-38
		FGFR1	5	2.4301E-11	ZWILCH	6	1.9536E-37
		FOSL1	5	4.2275E-09	ZNF460	6	2.7617E-36
		FRMD6	5	2.0781E-43	ZNF799	6	9.0052E-36
		GALNT3	5	1.0705E-09	TTTY3B	6	1.3485E-35
		GAS1	5	0.0205668	C8orf42	6	1.4534E-34
		GINS1	5	3.7888E-05	TEF	6	4.4741E-34
		GINS2	5	1.2702E-06	PCP4L1	6	4.9315E-33
		GJB3	5	1.0711E-06	SMCR7L	6	1.063E-32
		GLI3	5	2.4419E-16	MB	6	3.1442E-32
		GNAO1	5	2.8509E-07	AMY1A	6	6.3201E-32
		GSN	5	9.9467E-13	C20orf30	6	5.1127E-31

Concordant predictors for Both cohorts		Unique predictors for the TCGA- STAD- MutantTP53			Unique predictors for the TCGA- STAD- Wild TypeTP53		
Gene Name	P-value	Gene Name	Frequency	P-value	Gene Name	Frequency	P-value
		IL1R1	5	1.3111E-05	REV1	6	9.3997E-31
		ITGA2	5	0.04833419	PATL1	6	3.6803E-30
		ITGB4	5	0.04165069	ZMAT1	6	4.2243E-30
		JAM3	5	0.02368976	SMARCAL1	6	1.3066E-29
		ZNF192	5	0.00023229	ARL16	6	1.9579E-29
		ZNF195	5	2.1564E-13	DIAPH3	6	2.4625E-29
		ZNF638	5	9.2373E-05	TDG	6	5.0359E-29
		ATR	5	1.5518E-13	BCL2L14	6	2.7828E-28
		CHEK2	5	0.00033494	KRTAP4-9	6	3.5319E-28
		PERP	5	3.0984E-07	NCRNA00116	6	1.2253E-27
					BDH1	6	7.6676E-27
					GPD2	6	1.0258E-26
					HBZ	6	3.4249E-26

Concordant predictors for Both cohorts		Unique predictors for the TCGA- STAD- MutantTP53			Unique predictors for the TCGA- STAD- Wild TypeTP53		
Gene Name	P-value	Gene Name	Frequency	P-value	Gene Name	Frequency	P-value
					BARD1	6	5.1906E-26
					INHBA	6	1.0339E-25
					SLCO4A1	6	1.2793E-24
					KIF4B	6	3.1471E-24
					ZNF121	6	1.6912E-23
					LOC678655	6	3.8078E-23
					MXRA5	6	5.4158E-23
					PAR-SN	6	6.495E-23
					FOXF1	6	4.6164E-22
					RNFT1	6	1.8735E-21
					NOXO1	6	2.2536E-18
					APAF1	6	3.6193E-18
					PLEK2	6	5.6466E-18

Concordant predictors for Both cohorts		Unique predictors for the TCGA-STAD- MutantTP53			Unique predictors for the TCGA-STAD-Wild TypeTP53		
Gene Name	P-value	Gene Name	Frequency	P-value	Gene Name	Frequency	P-value
					WDR41	6	1.1471E-17
					TFAP2B	6	2.4733E-13
					MAGEE1	6	2.7762E-13
					HUS1	6	3.3757E-13
					SALL2	6	1.7545E-12
					LOC100128023	6	2.0282E-11
					GJB7	6	4.3009E-11
					ZNF626	6	7.6244E-11
					CD7	6	2.4226E-08
					AKAP1	6	2.394E-07
					MOBKL2B	6	1.4283E-06
					CHST10	6	2.2425E-05
					MAP2	6	1.3605E-19

Concordant predictors for Both cohorts		**Unique predictors for the TCGA-STAD- MutantTP53**			**Unique predictors for the TCGA-STAD-Wild TypeTP53**		
Gene Name	P-value	Gene Name	Frequency	P-value	Gene Name	Frequency	P-value
					SPAG17	6	7.3519E-45
					EPSTI1	6	0.03676313
					FUT8	6	2.0374E-18
					FOXP3	6	7.7557E-42
					PRRT3	6	0.00784276
					CLEC4A	6	0.00023229
					IQGAP3	5	2.301E-47
					RPL18A	5	9.0024E-47
					ITPR3	5	1.6704E-46
					C8orf46	5	3.5128E-45
					ZNF582	5	4.6471E-45
					C21orf33	5	4.9825E-45
					DCLRE1C	5	6.9453E-45

Concordant predictors for Both cohorts		Unique predictors for the TCGA- STAD- MutantTP53			Unique predictors for the TCGA- STAD- Wild TypeTP53		
Gene Name	P-value	Gene Name	Frequency	P-value	Gene Name	Frequency	P-value
					KIAA1804	5	8.9568E-45
					LCOR	5	3.8852E-44
					ATAD5	5	4.8367E-44
					C3orf10	5	6.3661E-44
					PLK2	5	1.2654E-43
					LOC100124692	5	2.0977E-43
					ERN2	5	6.8319E-43
					OVCH1	5	2.0867E-42
					LUZP4	5	3.2052E-42
					ARPP21	5	1.117E-40
					C19orf21	5	1.181E-40
					LGALS4	5	1.9382E-40
					CYB5D2	5	2.2031E-40

Concordant predictors for Both cohorts		Unique predictors for the TCGA- STAD- MutantTP53			Unique predictors for the TCGA- STAD- Wild TypeTP53		
Gene Name	P-value	Gene Name	Frequency	P-value	Gene Name	Frequency	P-value
					C3orf18	5	7.7417E-40
					PIAS1	5	2.3988E-39
					LTBR	5	1.0994E-38
					MAP7D1	5	1.1588E-37
					KIF14	5	2.0867E-37
					BAALC	5	4.7631E-37
					IGFBP3	5	6.241E-37
					C15orf40	5	1.5362E-36
					ACAP2	5	1.5529E-36
					PBX1	5	3.0335E-36
					TAB1	5	4.047E-36
					LGI3	5	4.1378E-36
					SIX6	5	5.384E-36

Concordant predictors for Both cohorts		Unique predictors for the TCGA-STAD- MutantTP53			Unique predictors for the TCGA-STAD-Wild TypeTP53		
Gene Name	P-value	Gene Name	Frequency	P-value	Gene Name	Frequency	P-value
					RNF11	5	7.7509E-36
					COL4A5	5	1.1462E-35
					CYBASC3	5	1.5208E-35
					FAM3C	5	6.2765E-35
					PMPCA	5	1.349E-34
					KCTD16	5	1.5905E-34
					APIP	5	1.901E-34
					MBNL2	5	2.2327E-34
					TJP3	5	4.4591E-34
					ZNF780B	5	6.6753E-34
					BTBD7	5	2.206E-33
					FAM168B	5	2.4893E-33
					FUT4	5	2.6216E-33

Concordant predictors for Both cohorts		**Unique predictors for the TCGA-STAD- MutantTP53**			**Unique predictors for the TCGA-STAD-Wild TypeTP53**		
Gene Name	P-value	Gene Name	Frequency	P-value	Gene Name	Frequency	P-value
					ATP5H	5	3.9558E-33
					SH2D2A	5	5.9131E-33
					FUCA2	5	1.6612E-32
					ADAM15	5	1.9313E-32
					PMS2	5	2.4436E-31
					KDM5A	5	6.7809E-31
					MYEOV	5	7.5172E-31
					GBAS	5	1.422E-30
					CRYGA	5	5.6845E-30
					ZNF219	5	6.3909E-30
					EHF	5	7.6045E-30
					ANLN	5	8.8111E-30
					MMP15	5	1.8708E-29

Concordant predictors for Both cohorts		Unique predictors for the TCGA-STAD- MutantTP53			Unique predictors for the TCGA-STAD-Wild TypeTP53		
Gene Name	P-value	Gene Name	Frequency	P-value	Gene Name	Frequency	P-value
					FLJ16779	5	6.0261E-29
					ROGDI	5	6.6792E-29
					ACCN4	5	7.207E-29
					KRT19	5	7.2982E-29
					DUSP4	5	1.2057E-28
					PHGR1	5	1.2531E-28
					C1orf141	5	1.5476E-28
					TNFSF11	5	1.7802E-28
					AIM1L	5	2.1567E-28
					CHMP4C	5	5.6088E-28
					GRHL2	5	6.2759E-28
					CA11	5	6.9769E-28
					UBE2MP1	5	1.9221E-27

Concordant predictors for Both cohorts		Unique predictors for the TCGA-STAD- MutantTP53			Unique predictors for the TCGA-STAD-Wild TypeTP53		
Gene Name	P-value	Gene Name	Frequency	P-value	Gene Name	Frequency	P-value
					IKBKE	5	8.0984E-27
					RRP15	5	1.1826E-26
					TPP2	5	2.0139E-26
					ASGR1	5	2.1417E-26
					MOCOS	5	2.6493E-26
					VNN1	5	4.7609E-26
					KIAA0495	5	6.6347E-26
					C14orf132	5	7.991E-26
					ST3GAL3	5	2.014E-25
					H2AFX	5	3.9296E-25
					CENPM	5	1.3868E-24
					IPMK	5	1.394E-24
					ZC3HAV1	5	1.5978E-24

Concordant predictors for Both cohorts		Unique predictors for the TCGA- STAD- MutantTP53			Unique predictors for the TCGA- STAD- Wild TypeTP53		
Gene Name	P-value	Gene Name	Frequency	P-value	Gene Name	Frequency	P-value
					CLDN12	5	1.7352E-24
					TNFRSF11A	5	5.756E-24
					DDX11	5	1.8311E-23
					OR4M1	5	2.1341E-22
					NAP1L2	5	4.2078E-22
					EPS8L1	5	1.2205E-21
					ZFP36L2	5	1.6409E-21
					CCDC30	5	2.2064E-20
					CHSY1	5	3.7852E-20
					TSNAXIP1	5	1.6335E-19
					CDH1	5	3.7988E-19
					ROM1	5	6.2864E-19
					RAET1L	5	3.4142E-18

Concordant predictors for Both cohorts		Unique predictors for the TCGA-STAD- MutantTP53			Unique predictors for the TCGA-STAD-Wild TypeTP53		
Gene Name	P-value	Gene Name	Frequency	P-value	Gene Name	Frequency	P-value
					C10orf81	5	2.1169E-17
					METT5D1	5	1.8681E-16
					HSPG2	5	2.0247E-16
					ISYNA1	5	2.4595E-16
					RB1	5	1.9103E-15
					NEFL	5	8.4488E-15
					SHC2	5	1.0789E-14
					ESRP2	5	1.277E-14
					SGK196	5	2.4175E-14
					SPATA19	5	2.6279E-14
					PIK3AP1	5	4.0195E-13
					DTNA	5	1.322E-12
					CSPP1	5	1.9677E-12

Concordant predictors for Both cohorts		Unique predictors for the TCGA- STAD- MutantTP53			Unique predictors for the TCGA- STAD- Wild TypeTP53		
Gene Name	P-value	Gene Name	Frequency	P-value	Gene Name	Frequency	P-value
					CYP3A4	5	4.9044E-12
					COL3A1	5	6.3362E-12
					ACAN	5	4.85E-10
					ADAMTS12	5	6.0752E-10
					IRF8	5	1.2387E-08
					STEAP1	5	8.9931E-08
					LOC642587	5	2.6381E-07
					CACNA2D2	5	3.1338E-07
					ZNF542	5	1.3236E-06
					SYCP2	5	3.3304E-05
					TGM2	5	0.00012235
					COX17	5	0.00187543
					SBNO2	5	0.03845241

Concordant predictors for Both cohorts		**Unique predictors for the TCGA-STAD- MutantTP53**			**Unique predictors for the TCGA-STAD-Wild TypeTP53**		
Gene Name	P-value	Gene Name	Frequency	P-value	Gene Name	Frequency	P-value
					PLSCR1	5	0.0481102
					LUM	5	0.04825685
					PDE4D	5	5.4974E-40
					AP3B2	5	1.6816E-20
					CASP4	5	8.0263E-40
					C6orf167	5	2.4238E-21
					SERP2	5	0.00031832
					CHRDL2	5	7.1464E-17
					C1orf96	5	6.5135E-11
					XRN1	5	5.5099E-26
					KCNMB1	5	7.0218E-06
					MCM4	5	1.2161E-18
					ZNF205	5	1.0797E-22

Concordant predictors for Both cohorts		**Unique predictors for the TCGA- STAD- MutantTP53**			**Unique predictors for the TCGA- STAD- Wild TypeTP53**		
Gene Name	P-value	Gene Name	Frequency	P-value	Gene Name	Frequency	P-value
					MYOCD	5	0.00033494
					PPP1R3C	5	2.0635E-26
					SNAP25	5	3.6648E-39
					PCDHGA1	5	0.00016435
					TFRC	5	1.3947E-05
					PRDM9	5	0.00061516
					FNDC1	5	0.00135475
					ASB2	5	3.0984E-07
					CD3G	5	0.00093811
					GVIN1	5	0.00270394
					CCL18	5	1.8459E-19
					CTNND2	5	1.942E-10
					CRISPLD2	5	7.1377E-12

Concordant predictors for Both cohorts		**Unique predictors for the TCGA-STAD- MutantTP53**			**Unique predictors for the TCGA-STAD-Wild TypeTP53**		
Gene Name	P-value	Gene Name	Frequency	P-value	Gene Name	Frequency	P-value
					OR5T2	5	9.2373E-05
					ATR	5	1.5518E-13

TCGA-PAAD Project

Concordant predictors for Both cohorts		Unique predictors for the TCGAPAAD-MutantTP53			Unique predictors for the TCGAPAAD-Wild TypeTP53		
Gene Name	P-value	Gene Name	Frequency	P-value	Gene Name	Frequency	P-value
E2F1	0.004737	EBF1	8	0.028625	S100A16	12	8.69E-05
KIAA0101	0.000252	KIF18B	8	0.000332	EFNA4	9	0.000472
ORC6L	6.17E-05	OIP5	8	1.45E-06	OSBPL3	9	8.79E-05
REV3L	0.009449	SPAG5	8	0.001426	S100A11	9	1.45E-05
TPX2	1.92E-05	BUB1	7	5.42E-05	TMEM92	9	0.000186
ZWINT	7.26E-05	C1orf135	7	1.11E-05	ALPK1	8	0.012528
CDC20	0.000502	CDK1	7	1.62E-05	ANXA11	8	0.000181
CDC45	0.003119	CDKN3	7	0.003292	C19orf33	8	3.51E-05
CDC6	0.001088	CENPN	7	0.005529	FNDC3A	8	0.000577

Concordant predictors for Both cohorts		**Unique predictors for the TCGAPAAD-MutantTP53**			**Unique predictors for the TCGAPAAD-Wild TypeTP53**		
Gene Name	P-value	Gene Name	Frequency	P-value	Gene Name	Frequency	P-value
FAM72B	0.000178	DTYMK	7	0.000416	KLF5	8	0.003597
ITGB4	0.000397	FAM72D	7	0.000171	PLEK2	8	4.55E-06
MYBL2	0.000147	FAM83H	7	0.000351	S100A6	8	0.000332
UBE2C	1.6E-05	GTSE1	7	0.006853	TRIM16	8	0.004138
ANLN	1.22E-05	MAD2L1	7	0.002105	ATP2B1	7	0.000735
ASF1B	0.000331	MCM10	7	4.33E-05	C6orf132	7	0.000707
AURKA	0.001103	MCM4	7	0.001224	CARD6	7	0.034516
AURKB	0.036029	PKMYT1	7	0.000558	CMTM7	7	0.000202
BIRC5	0.001604	POLQ	7	0.002132	DEPDC1B	7	0.027248
C17orf53	0.00086	RACGAP1	7	0.00189	E2F8	7	0.002327
DTL	0.009477	SMAD5	7	0.001188	ECT2	7	0.00031
EPR1	0.001103	SOCS2	7	8.27E-06	ERBB2	7	8.99E-05

Concordant predictors for Both cohorts		Unique predictors for the TCGAPAAD-MutantTP53			Unique predictors for the TCGAPAAD-Wild TypeTP53		
Gene Name	P-value	Gene Name	Frequency	P-value	Gene Name	Frequency	P-value
EXO1	0.004438	TACC3	7	0.011886	FHL2	7	0.000319
HJURP	2.25E-05	TNRC6C	7	1.75E-05	FRRS1	7	0.003143
KIF15	0.00025	UBE2T	7	0.003244	GBP2	7	0.001526
KIF2C	0.003119	AP1S3	6	4.4E-06	CLIC1	5	8.99E-05
LAMB3	0.001088	C15orf42	6	0.001074	CMTM7	7	0.000202
MELK	0.000178	CCNB1	6	0.000948	CPEB4	6	0.030645
NCAPG	0.000397	CDCA4	6	6.38E-06	CREBL2	5	0.011294
POC1A	0.000147	E2F7	6	0.000199	CTNNA1	5	2.95E-05
RAD51	1.6E-05	FAM54A	6	0.031451	DEPDC1B	7	0.027248
RAD54L	1.22E-05	FANCA	6	0.000267	E2F8	7	0.002327
SKA1	0.000331	FANCB	6	0.000946	ECM1	5	0.004254
BUB1B	0.001103	FANCD2	6	0.016025	ECT2	7	0.00031

Concordant predictors for Both cohorts		Unique predictors for the TCGAPAAD-MutantTP53			Unique predictors for the TCGAPAAD-Wild TypeTP53		
Gene Name	P-value	Gene Name	Frequency	P-value	Gene Name	Frequency	P-value
C11orf82	0.036029	GHR	6	0.024427	EFNA4	9	0.000472
CCNA2	0.001604	GIMAP7	6	0.010649	EPS8	5	0.002502
CCNB2	0.00086	GNG2	6	0.034834	EPS8L1	6	6.29E-05
CDC25C	0.004438	GPR56	6	0.016224	ERBB2	7	8.99E-05
CDCA5	2.25E-05	HERC1	6	0.002092	ESPL1	6	0.011441
CENPE	0.00025	KRT19	6	8.71E-05	ETV6	5	0.006154
CEP55	0.003119	MCM2	6	0.000184	FAM83A	5	1.1E-08
CKAP2L	0.001088	MLF1IP	6	0.005946	FCGR2C	5	0.015442
CKS2	0.000178	PLK4	6	0.003457	FGD6	5	1.32E-05
DLGAP5	0.000397	POLE2	6	0.012267	FHL2	7	0.000319
EME1	0.000147	PPP3CB	6	0.000635	FICD	5	0.000852
ERCC6L	0.000252	RECQL4	6	0.014789	FLJ23867	6	0.00065
FAM64A	6.17E-05	RFC4	6	0.001206	FMO1	6	0.014526
GINS1	0.009449	RNASEH2A	6	0.013048	FNDC3A	8	0.000577
KIF18A	1.92E-05	SP3	6	0.032765	FOXQ1	6	0.001168

Concordant predictors for Both cohorts		Unique predictors for the TCGAPAAD-MutantTP53			Unique predictors for the TCGAPAAD-Wild TypeTP53		
Gene Name	P-value	Gene Name	Frequency	P-value	Gene Name	Frequency	P-value
KIF23	7.26E-05	TRIP13	6	0.000958	FRRS1	7	0.003143
KIFC1	0.000502	TTK	6	0.00014	FSTL1	5	0.009385
KLHL8	0.003119	USP38	6	0.025547	FUT3	5	0.009743
LAMC2	0.001088	WDR62	6	0.001506	GALNT6	5	0.049224
MET	0.000178	ABCD2	5	0.009565	GBP2	7	0.001526
MND1	0.000397	APC	5	0.015096	GFPT2	5	0.015807
NCAPH	0.000147	ARHGAP11A	5	0.000408	GPRC5A	7	8.26E-05
NEK2	1.6E-05	BCL2L1	5	0.000569	GPX8	5	0.000268
NUF2	1.22E-05	C9orf140	5	0.001455	GRHL2	7	0.002233
NUSAP1	0.000331	CCND1	5	0.000353	GUK1	5	0.002875
SGOL1	0.001103	CCT5	5	0.001123	HAGH	5	0.013516
SHCBP1	0.036029	CDCP1	5	2.37E-06	HDHD2	6	1.24E-05
TK1	0.001604	DICER1	5	0.014328	HMGA2	7	0.000101
TOP2A	0.00086	DVL1	5	0.001809	HMMR	5	0.000212
ASPM	0.009477	ELMO1	5	0.00029	IGF2BP2	5	0.00034

Concordant predictors for Both cohorts		**Unique predictors for the TCGAPAAD-MutantTP53**			**Unique predictors for the TCGAPAAD-Wild TypeTP53**		
Gene Name	P-value	Gene Name	Frequency	P-value	Gene Name	Frequency	P-value
C16orf75	0.001103	FAM72A	5	0.001414	IGFBP7	5	0.045159
C1orf106	0.004438	FBN1	5	0.029747	IL18	5	3.18E-05
C1orf112	2.25E-05	FBXL15	5	0.025065	IQGAP3	5	0.000449
CDCA8	0.00025	GJB3	5	4.12E-05	IRF2BP1	5	0.000525
CENPA	0.000178	GTF3C3	5	0.029453	ITGB6	6	4.04E-05
CENPF	0.000397	H2AFZ	5	0.00589	ITPR3	6	0.000229
CENPI	0.000147	HELLS	5	0.001426	KCNN4	5	2.47E-06
CENPK	0.000252	ITGA2	5	2.96E-05	KLF5	8	0.003597
CENPM	6.17E-05	KCTD12	5	0.006755	KRT15	5	0.000293
COL14A1	0.009449	KIF14	5	0.000745	KYNU	5	0.000503
DEPDC1	1.92E-05	KIF22	5	0.023699	LGALS3	7	0.000264
EPHA2	7.26E-05	KRT7	5	6.62E-06	LGALS9	6	0.001236
EZH2	0.000502	LAMA3	5	2.45E-06	LPAR5	5	0.043872
FANCI	0.003119	MAP3K2	5	0.045311	LRRC8E	7	0.001481
FOXM1	0.001088	METTL7A	5	0.000154	LY6E	5	8.36E-07

Concordant predictors for Both cohorts		**Unique predictors for the TCGAPAAD-MutantTP53**			**Unique predictors for the TCGAPAAD-Wild TypeTP53**		
Gene Name	P-value	Gene Name	Frequency	P-value	Gene Name	Frequency	P-value
KIF11	0.000178	MFSD10	5	0.007827	MAPK10	5	0.017987
KIF20A	0.000397	MZF1	5	0.003965	MBOAT1	5	0.016538
KIF4A	0.001103	NCAPG2	5	0.010684	MFSD2B	7	5.05E-05
MKI67	0.036029	NEIL3	5	0.001942	MLKL	6	0.001386
NDC80	0.001604	ORC1L	5	0.026493	MST1R	7	0.000786
PBK	0.00086	PLK1	5	5.58E-05	MSX2	6	0.001264
PRC1	0.009477	RANBP1	5	0.033819	MVP	6	0.001289
PTTG1	0.001103	RHOD	5	3.46E-06	MYD88	5	0.000195
RRM2	0.004438	SDC4	5	3.46E-06	MYEOV	6	0.000244
SKA3	2.25E-05	SEPP1	5	7.6E-05	MYOF	5	3.08E-05
SMG1	0.00025	SGOL2	5	0.023939	NDE1	6	2.52E-05
TGFA	0.000178	SH2D3A	5	0.001877	NFKB2	6	0.028741
TROAP	0.000397	SHE	5	0.047737	NRM	7	6.66E-05
UHRF1	0.000147	SMC4	5	0.000562	NXF2B	6	0.000766
		SPC24	5	9.97E-05	OSBPL3	9	8.79E-05

Concordant predictors for Both cohorts		**Unique predictors for the TCGAPAAD-MutantTP53**			**Unique predictors for the TCGAPAAD-Wild TypeTP53**		
Gene Name	P-value	Gene Name	Frequency	P-value	Gene Name	Frequency	P-value
		SSH3	5	4.65E-05	P2RY2	5	7.61E-07
		TADA1	5	0.009931	PAIP2	6	0.012369
		TANK	5	0.030459	PDP1	6	0.000473
		TRAIP	5	0.003182	PDZK1IP1	5	0.004897
		TRIM23	5	0.03762	PEG3	5	0.045423
		VCAN	5	0.0004	PLCD3	5	0.000142
		ZFX	5	0.049586	PLEK2	8	4.55E-06
		AIFM2		0.026586	PLEKHN1	6	2.95E-06
		APAF1		0.039073	POLD4	5	0.000667
		BAX		0.024581	PPAP2C	5	0.013835
		BID		0.003637	PPFIA3	5	0.037262
		CASP8		0.003342	PPP1CA	7	0.000652
		CCNE1		0.000336	PRMT10	5	0.001055
		CCNG1		0.00275	PSMB8	5	0.001593
		CDK2		0.002537	PTMA	5	1.56E-06

Concordant predictors for Both cohorts		Unique predictors for the TCGAPAAD-MutantTP53			Unique predictors for the TCGAPAAD-Wild TypeTP53		
Gene Name	P-value	Gene Name	Frequency	P-value	Gene Name	Frequency	P-value
		CDK6		0.001093	PTPN12	7	0.000322
		CDKN2A		3.67E-05	PVRL4	7	0.000453
		CHEK1		0.004123	QDPR	6	0.002131
		CHEK2		0.020645	RAD51AP1	6	0.008212
		DDB2		9.45E-05	RALB	5	8.66E-08
		GADD45A		0.006611	RBM23	5	0.002261
		GADD45G		0.001576	RELA	5	0.000146
		IGFBP3		9.99E-06	RHBDF1	5	0.00019
		MDM2		2.1E-05	RHBDL2	5	0.000365
		MDM4		0.001537	RNF149	5	1.6E-06
		PERP		0.000134	RTN1	6	0.02971
		PMAIP1		0.014088	S100A11	9	1.45E-05

Concordant predictors for Both cohorts		Unique predictors for the TCGAPAAD-MutantTP53			Unique predictors for the TCGAPAAD-Wild TypeTP53		
Gene Name	P-value	Gene Name	Frequency	P-value	Gene Name	Frequency	P-value
		PPM1D		8.11E-07	S100A14	7	0.000254
		RCHY1		0.002448	S100A16	12	8.69E-05
		RPRM		0.012027	S100A6	8	0.000332
		RRM2B		0.006333	SERPINB5	5	0.0001
		SERPINB5		0.0001	SEZ6	7	0.010111
		SERPINE1		0.030528	SF3B1	6	0.024334
		SESN1		9.22E-05	SFN	7	5.1E-05
		SFN		5.1E-05	SIK3	5	0.00227
		SHISA5		0.003162	SLC26A11	6	1.24E-05
		SIAH1		0.006535	SLC39A1	6	7.61E-05
		STEAP3		4.6E-05	SMARCA2	5	0.000184
		THBS1		0.036411	SNAPC4	5	0.001722

Concordant predictors for Both cohorts		Unique predictors for the TCGAPAAD-MutantTP53			Unique predictors for the TCGAPAAD-Wild TypeTP53		
Gene Name	P-value	Gene Name	Frequency	P-value	Gene Name	Frequency	P-value
		TNFRSF10B		0.00098	SPPL2B	5	0.004074
		TP73		0.000133	STIL	5	0.00265
		CD82		0.007478	TACSTD2	5	0.000262
					TM4SF1	5	2.22E-05
					TMC7	7	5.38E-05
					TMEM92	9	0.000186
					TMOD3	5	5.32E-05
					TMPRSS4	5	5.26E-05
					TMSB10	7	2.43E-05
					TNFRSF10B	6	0.00098
					TNS4	6	3.87E-07
					TPBG	6	0.000283

Concordant predictors for Both cohorts		**Unique predictors for the TCGAPAAD-MutantTP53**			**Unique predictors for the TCGAPAAD-Wild TypeTP53**		
Gene Name	P-value	Gene Name	Frequency	P-value	Gene Name	Frequency	P-value
					TPM4	6	9.5E-05
					TRIM16	8	0.004138
					TRIM34	6	0.02517
					TRIP10	5	1.61E-05
					TSKU	5	6.19E-05
					TSPYL4	5	0.040254
					TUBA1C	5	0.000449
					TWF2	5	0.000229
					VAMP8	5	0.000515
					ZFP36L1	6	6.53E-05
					ZNF267	5	0.007883
					ZNF441	6	0.006687

Concordant predictors for Both cohorts		Unique predictors for the TCGAPAAD-MutantTP53			Unique predictors for the TCGAPAAD-Wild TypeTP53		
Gene Name	P-value	Gene Name	Frequency	P-value	Gene Name	Frequency	P-value
					ZNF540	6	0.001851
					ZNF625	5	0.030366
					ZNF709	6	0.010831
					SERPINE1		0.030528
					SESN1		9.22E-05
					SHISA5		0.003162
					SIAH1		0.006535
					STEAP3		4.6E-05
					THBS1		0.036411
					TP73		0.000133
					CD82		0.007478
					AIFM2		0.026586

Concordant predictors for Both cohorts		**Unique predictors for the TCGAPAAD-MutantTP53**			**Unique predictors for the TCGAPAAD-Wild TypeTP53**		
Gene Name	P-value	Gene Name	Frequency	P-value	Gene Name	Frequency	P-value
					APAF1		0.039073
					BAX		0.024581
					BCL2L1		0.000569
					BID		0.003637
					CASP8		0.003342
					CCNB1		0.000948
					CCND1		0.000353
					CCNE1		0.000336
					CCNG1		0.00275
					CDK1		1.62E-05
					CDK2		0.002537
					CDK6		0.001093

Concordant predictors for Both cohorts		Unique predictors for the TCGAPAAD-MutantTP53			Unique predictors for the TCGAPAAD-Wild TypeTP53		
Gene Name	P-value	Gene Name	Frequency	P-value	Gene Name	Frequency	P-value
					CDKN2A		3.67E-05
					CHEK1		0.004123
					CHEK2		0.020645
					DDB2		9.45E-05
					GADD45A		0.006611
					GADD45G		0.001576
					GTSE1		0.006853
					IGFBP3		9.99E-06
					MDM2		2.1E-05
					MDM4		0.001537
					PERP		0.000134
					PMAIP1		0.014088

Concordant predictors for Both cohorts		Unique predictors for the TCGAPAAD-MutantTP53			Unique predictors for the TCGAPAAD-Wild TypeTP53		
Gene Name	P-value	Gene Name	Frequency	P-value	Gene Name	Frequency	P-value
					PPM1D		8.11E-07
					RCHY1		0.002448
					RPRM		0.012027
					RRM2B		0.006333