Protein Networks in Colorectal Cancer Cells

Submitted by

David Chisanga

BSc. CMP.Sc, MSc.ITM

A thesis submitted in total fulfilment of the requirements for the degree of

Doctor of Philosophy

Department of Computer Science and Information Technology

School of Engineering and Mathematical Sciences

College of Science, Health and Engineering

La Trobe University

Bundoora, Victoria, 3086, Australia

December 2017

Table of contents

List of figures

List of tables

Abstract

Colorectal cancer (CRC) is the third most common form of cancer and has one of the highest rates of morbidity and mortality in the world. To understand the pathogenesis, progression and metastasis of CRC, biomedical researchers with the help of new high-throughput data collection techniques such as mass spectrometry (MS) and next-generation sequencing (NGS) coupled with innovative experimental strategies perform global analyses of entire whole-genomes, transcriptomes and proteomes. These new developments have in turn led to a surge in both qualitative and quantitative omics data which now pose analytical challenges for biologists on how to infer clinically relevant insights on the disease. Nonetheless, cancer is known as a disease of the pathways and as such, understanding the structure, dynamics and interactions of biological molecules such as proteins in protein-protein interactions (PPI) and the role of extracellular vesicles such as exosomes in cancer can help us understand the intricacies involved in cancer.

The main objective of this thesis is, therefore, to develop bioinformatics tools and resources for the collation and analysis of CRC related omics data as well as the inference of CRC biomarkers from the dynamic changes that take place in PPIs. Thus, we developed the Colorectal Cancer Atlas, a web-based online platform that collates and integrates multiple CRC-related omics data. To understand the dynamic changes that take place in PPIs in CRC cells, we integrated the collated heterogenous omics datasets by applying network theory methods and a machine learning approach and inferred new as well as known biomarkers that can be used to study the pathogenesis and progression of CRC. In addition, using an integrated bioinformatics approach that combines network theory and physical coherence, we also identified NEDD4 and STAMBP as novel regulators of exosome biogenesis.

Statement of authorship

This thesis includes work by the author that has been published or accepted for publication as described in the text. Except where reference is made in the text of the thesis, this thesis contains no material published elsewhere or extracted in whole or in part from a thesis accepted for the award of any other degree or diploma.

No other person's work has been used without due acknowledgement in the main text of this thesis. This thesis has not been submitted for the award of any degree or diploma in any other tertiary institution.

David Chisanga

4th December 2017

Acknowledgements

First and foremost, I would like to thank God, through his son, Jesus Christ for giving me the strength and the will to keep going through the ups and downs during the candidature.

I would further like to thank Dr Suresh Mathivanan for giving me the opportunity to work in his laboratory as well as the opportunities to work on various projects that enabled me to gain valuable skillsets. I will also be forever grateful to Dr Mathivanan for the financial support he provided at the end of my scholarship as well as the constant advice and support which was helpful in completing my PhD studies. I would also like to express my sincere appreciation to him for his efforts in ensuring that I had a smooth transition to a new supervisor after the departure of my first supervisor, Dr Johnson I Agbinya.

Further appreciation goes to my principal supervisor, Dr Naveen Chilamkurti for his constant support and advice. I would also like to thank him for his willingness to take on the supervision of my candidature. His valuable insight into networks and networking proved very useful in my research. I would also like to acknowledge my co-supervisor, Dr Shivakumar Keerthikumar, for first agreeing to be my co-supervisor as well as the bioinformatics and data analytics insight he rendered. I am also thankful for his advice and his invaluable contributions to my research projects. I would also like to thank Dr Agus Salim and the Research Progressive Panel (RPP) chair, Dr Ben Soh. Special thanks also go to my laboratory mates for their help and support: Mohashin Pathan, Sushma Anand, Monisha Samuel, Kening Zhao, Dr Lahiru Gangonda, Dr Hina Kalra, Ishara Atukorala, Michael Liem, Nidhi Abraham, Pamali Fonseka, and Stephanie Boukouris. I would also like to extend my appreciation to Dr Johnson I Agbinya in his role as my first principal supervisor and for the valuable advice he provided. I will forever be indebted to him for helping me settle in into life at La Trobe University. I would also like to extend my sincere appreciation to my friends and colleagues who helped me in any way during my research such as Abebe Diro for his assistance on machine learning and Clarence Leon, Dr Hoang Nguyen, and Dr Hoa Doan Thanh for their help and guidance as I settled in into the Department of Electronic Engineering in my first year of candidature.

I would also like to acknowledge and thank La Trobe University for providing me with the scholarships (LTUFFRS and LTUPRS) and the opportunity to pursue my PhD studies.

Lastly, I would like to express my heartfelt appreciation to my wife, Mubanga and my daughter Joy, firstly, for enduring the lonely nights when I was working nights at the office and secondly, for my wife's unwavering support and prayers, and for the love and encouragement she gave me in the lowest moments of my candidature. I am forever indebted to her. I would also like to say a big thank you to Joy whose smiles would cheer me up even when things were not going as expected. I am also forever grateful to my family; to Dad, thank you for believing in us (your children) and giving us the free will to explore our interests and what we were passionate about. To mum, brothers (Kasongo, Hopkins, Kevin) and sisters (Chisala, Leticia), I am thankful for your constant encouragement. I would also like to thank my sisters (Bana Mambwe, Bana Malifwa and Ba Musakanya) and brother-in-law (Major Bwalya) whose encouragement and prayers were a reminder of what family means. I would also like to give a special mention to Bana Mambwe for coming down to help us with Joy and to General Mulenga for the support during her travel and for sponsoring her trip.

Abbreviations

AB	Apoptotic Bodies
AI	Artificial intelligence
BC	Betweenness Centrality
BIND	Biomolecular Interaction Network Database
BioGRID	Biological General Repository for Interaction Database
BLAST	Basic Local Alignment Search Tool
CC	Closeness Centrality
CIMP	CpG island methylator phenotype
CIN	Chromosomal Instability
COSMIC	Catalogue of Somatic Mutations in Cancer
CRC	Colorectal Cancer
CRISPR	Clustered Regularly Interspaced Short Palindromic Repeats
D3JS	Data-Driven Documents
DC	Degree Centrality
DIP	Database of Interacting Proteins
DNA	Deoxyribonucleic Acid
EMT	Epithelial-to-Mesenchymal Transition
ESCRT	Endosomal Sorting Complex Required for Transport
EV	Extracellular Vesicle
FACS	Fluorescence-Activated Cell Sorting
FDR	False Discovery Rate
FunRich	Functional Enrichment Tool
GBA	Guilty by Association
GDC	Genomic Data Commons
GEO	Gene Expression Omnibus
GLM	Generalized Linear Model
GO	Gene Ontology
GWAS	Genome-Wide Association Study
HIPPIE	Human Integrated Protein-Protein Interaction rEference
HPRD	Human Protein Reference Database
HTML	Hypertext Mark-up Language

HTP	High-throughput
ILV	Intraluminal Vesicle
IntAct	IntAct molecular interaction database
LAC	Local Area Connectivity
LS	Lynch syndrome
LTP	Low-throughput
MGF	Mascot Generic File Format
MINT	Molecular Interaction Database
mRNA	Messenger Ribonucleic Acid
MS	Mass Spectrometry
MSI	Microsatellite Instability
MVBs	Multivesicular Bodies
NGS	Next Generation Sequencing
NSAF	Normalized Spectral Abundance Factor
PIE	Physical Interaction Enrichment
PPI	Protein-Protein Interaction
PTMs	Post-translational modifications
RNA	Ribonucleic Acid
RNA-Seq	RNA Sequencing
scRNA-Seq	Single Cell RNA Sequencing
SMVs	Shedding Microvesicles
STRING	Search Tool for Recurring Instances of Neighbouring Genes/Proteins
TCGA	The Cancer Genome Atlas
WCRF	World Cancer Research Fund International
Y2H	Yeast-2-Hybrid

Publications and conference papers

First author refereed publications;

D. Chisanga, S. Keerthikumar, and N. Chilamkurti, "Network tools for the analysis of proteomic data," in *Proteome Bioinformatics*, S. Mathivanan, Ed.: Springer, 2017. (Chapter 2).

D. Chisanga, S. Keerthikumar, M. Pathan, D. Ariyaratne, H. Kalra, S. Boukouris, N. A. Mathew, H. A. Saffar, L. Gangoda, C.-S. Ang, O. M. Sieber, J. M. Mariadason, R. Dasgupta, N. Chilamkurti, and S. Mathivanan, "Colorectal cancer atlas: An integrative resource for genomic and proteomic annotations from colorectal cancer cell lines and tissues," *Nucleic Acids Research,* vol. 44, no. D1, pp. D969-D974, January 4, 2016. (Chapter 3).

Submitted first author refereed journal publications

D. Chisanga, S. Keerthikumar, N. Chilamkurti, and S. Mathivanan, “Perturbation of protein-protein interaction network based on APC mutations in Colorectal Cancer,”. Submitted to the journal *Scientific Reports,* December 2017 (Chapter 4)

D. Chisanga, S. Keerthikumar, S. Mathivanan, and N. Chilamkurti, “Integration of heterogeneous ‘Omics’ Data Using Semi-Supervised Network Labelling to Identify Essential Genes in Colorectal Cancer,” Journal of *Computers and Electronic Engineering*, 2018. (Chapter 5)

The candidate also contributed to the following journal publications during the PhD candidature

K. Zhao, M. Bleackley, **D. Chisanga**, L. Gangoda, M. Liem, H. Kalra, HA. Saffar, S. Keerthikumar, C. Ang, C. Adda, L. Jiang, K. Yap, I. Poon, P. Lock, V. Bulone, M. Anderson and S. Mathivanan, “Novel cell wall remodelling functions of extracellular vesicles secreted by Saccharomyces cerevisiae,” submitted to the journal *Scientific Reports,* November 2016

M. Samuel, L. Gangoda, S. Keerthikumar, A. Spurling, C. Ang, C. Vennin, MC. Lucas, L. Cheng, D. Herrmann, M. Pathan, **D. Chisanga**, SC. Warren, P. Fonseka, N. Abraham, S. Anand, S. Boukouris, CG. Adda, L. Jiang, TM. Shekhar, N. Baschuk, CJ. Hawkins, AJ. Johnston, NJ. Hoogenraad, IK. Poon, AF. Hill, M. Jois, P. Timpson, BS. Parker and S. Mathivanan, “Oral administration of bovine milk-derived exosomes reduce primary tumor burden but accelerate cancer metastases,” submitted to the Journal *Science, November 2016.*

M. Samuel, **D. Chisanga**, M. Liem, S. Keerthikumar, S. Anand, C-S. Ang, CG. Adda, E . Versteegen, M. Jois and S. Mathivanan, “Bovine milk-derived exosomes from colostrum are enriched with proteins implicated in immune response and growth,” *Scientific Reports*, 2017, vol. 7.

M. Pathan, S. Keerthikumar, **D. Chisanga**, R. Alessandro, C.-S. Ang, P. Askenase, A. O. Batagov, A. Benito-Martin, G. Camussi, A. Clayton, F. Collino, D. Di Vizio, J. M. Falcon-Perez, P. Fonseca, P. Fonseka, S. Fontana, Y. S. Gho, A. Hendrix, E. N.-t. Hoen, N. Iraci, K. Kastaniegaard, T. Kislinger, J. Kowal, I. V. Kurochkin, T. Leonardi, Y. Liang, A. Llorente, T. R. Lunavat, S. Maji, F. Monteleone, A. Øverbye, T. Panaretakis, T. Patel, H. Peinado, S. Pluchino, S. Principe, G. Ronquist, F. Royo, S. Sahoo, C. Spinelli, A. Stensballe, C. Théry, M. J. C. van Herwijnen, M. Wauben, J. L. Welton, K. Zhao, and S. Mathivanan, "A novel community driven software for functional enrichment analysis of extracellular vesicles data," *Journal of Extracellular Vesicles,* vol. 6, no. 1, p. 1321455, 2017/01/01 2017.

S. Keerthikumar, **D. Chisanga**, D. Ariyaratne, H. Al Saffar, S. Anand, K. Zhao,M. Samuel, M. Pathan, M. Jois, N. Chilamkurti, L. Gangoda and S. Mathivanan, “ExoCarta: A Web-Based Compendium of Exosomal Cargo,” *Journal of Molecular Biology*, 2016, vol. 428, no. 4, pp. 688-692.

This thesis also includes data to be submitted to a peer-reviewed journal

D. Chisanga, A. Sushma, et al, “Physical coherence and network analysis reveals NEDD4 as novel regulator of exosomes biogenesis,” manuscript in preparation (Chapter 7).

Refereed conference proceedings

D. Chisanga, A. Sushma, et al, “Physical coherence and network analysis reveals NEDD4 as novel regulator of exosomes biogenesis,” Australian Bioinformatics and Computational Biology Society, Adelaide, Australia. 2017

D. Chisanga, M. Pathan, et al, “Colorectal Cancer Atlas and FunRich: Discovery tools for integrated 'omics' data analysis,” Australian Bioinformatics and Computational Biology Society, Queensland University of Technology (QUT - GARDENS POINT CAMPUS), Brisbane, Australia. 2016

Chapter 1

General introduction

Cancer is a disease which results from the dysregulation or hyperactivity in the network of intracellular and extracellular signalling cascades, leading to the abnormal and uncontrolled growth of cells [1]. It is one of the leading causes of death around the world and is one of the most significant health challenges facing humanity today. For instance, in 2012, over 14 million people worldwide were diagnosed with cancer, and in 2015, there were over eight million cancer-related deaths worldwide, making it the second leading cause of death [2, 3]. Moreover, projections are that over the next twenty years, new cancer cases are expected to rise by more than seventy percent (70%) [4]. The disease is a significant cause of pain and distress, not only to patients but also to those around them, leading to loss in terms of a productive workforce and stress on the healthcare systems of nations across the globe. Of the many types of cancer, colorectal cancer (CRC) is the third most common form of cancer and has one of the highest rates of morbidity and mortality in the world if not treated in time.

There are several ongoing efforts from the scientific community around the world to understand the genesis, progression and metastasis of CRC as well as the development of solutions to tackle this disease. However, despite the incredible progress that has been achieved so far in developing solutions that can help contain the disease, drugs, if available, are usually too expensive for many patients to afford and some of the therapies available become ineffective as patients develop resistance. Predicting a patient's response to therapy to develop individualised treatments remains one of the most significant challenges given that most of the cancer drugs available are only able to work on a fraction of the patients [2, 5].

To understand cancer, researchers study how biological systems function and how their functionality is modified during cancer progression. Biological systems function through a complex network of cellular processes in which various molecules such as proteins, metabolites and ribonucleic acids (RNAs) take part in a meticulously regulated manner. For instance, it has been shown by several researchers that cancer is the result of the dysregulation of pathways [1, 6]. This view is reaffirmed by Zuckerkandl and Pauling [7] who describe life as a relationship between molecules, and not a characteristic of a single molecule, and as such a breakdown in this relationship may lead to pathological conditions. Among these molecules are proteins which form the core of various cellular events, and their altered behaviours have been implicated in disease pathologies such as cancer. Understanding the structure, dynamics and interactions of proteins is one of the essential areas of research in the biomedical arena. In recent years, advancements in high-throughput data collection techniques such as mass spectrometry (MS), next generation sequencing (NGS) and single-cell RNA-Seq have enabled the study of proteomes, genomes and transcriptomes on a large-scale. Coupled with some of the latest experimental strategies as well as advances in computational tools and methods, high-throughput techniques now support the global study of cellular genomes and proteomes. These new developments have led to an increase in both qualitative and quantitative proteomics, genomics and transcriptomics data which now poses analytical challenges for biologists on how to infer clinically relevant insights on diseases such as cancer. It has, therefore, become impractical to map the vast datasets to biological processes using traditional methods, and the need for computer-aided data analytics methods is on the rise.

Networks offer novel ways by which complex biological datasets can be analysed to study the interplay of proteins. The use of networks to model protein interactions in human disease provides us with a simplified representation of the cell’s intricate wiring whose analysis can provide us with clues to understanding a disease [8]. However, existing interactome maps tend to be biased towards proteins implicated in diseases [9] and available tools for exploring these interactome networks in diseases are limited [10]. The application of traditional statistical tools on the supposition that quantities have a normal distribution or those representing various activity patterns are independent variables renders current tools ineffective. In addition, there are two types of protein networks, stable and transient networks. Most existing network-based analysis tools and methods overlook the dynamism of protein interactions in transient protein networks and focus more on static networks. Cancer is a heterogeneous disease whereby individuals with the same type of cancer can have different forms of the same disease. Networks in such cases can be used to develop personalised profiles of such individuals. However, in most of the literature, protein networks are usually studied as static networks, even when data sets are collected at different time points, at different conditions and with different technologies.

The analysis of biological interactomes using static networks does not address several factors which need to be taken into account. Such factors include biological functions, being time-sensitive, proteins and the fact that the networks they form do not always exist at the same time. In addition, biological networks are dynamic in nature, a single protein can serve multiple functions and at the same time can interact with proteins that function entirely differently from its own. A static network will therefore not account for the spatial and temporal aspects of biological interactomes and may lead to the inaccurate representation of the dynamism that is characteristic of biological networks. To therefore correctly analyse proteomics data in heterogeneous diseases such as cancer, there is a need to develop computational methods and tools that can encompass the temporal aspects underlying such diseases.

1.1. Aims

The primary aim of this thesis is to develop bioinformatics tools and resources based on network theory for the analysis of CRC related omics data. This aim was further split into three sub-aims as follows:

i. Develop an integrated repository for CRC-related omics data
ii. Integration of omics data using network theory and machine learning-based methods to identify essential genes in CRC samples
iii. Identify novel proteins that regulate exosome biogenesis using bioinformatics

1.2. Thesis overview

The thesis is structured as follows;

- Chapter 2: *Background and related work.* This chapter gives a general overview of the current knowledge on CRC. The chapter further discusses the current knowledge of the roles of PPIs in cancer together with the recent tools and methods that are available for the network analysis of PPIs. In addition, the chapter discusses some of the computational tools that have been developed and applied in the network analysis of omics data to infer essential genes in cancer. Finally, the role of exosomes in physiological and pathological states with a view to understanding exosome biogenesis are discussed.

- Chapter 3: *Colorectal Cancer Atlas: An integrative resource for genomic and proteomic annotations from colorectal cancer cell lines and tissue*. This chapter has been peer-reviewed and published in the journal *Nucleic Acid Research* [11] and is presented here as a manuscript. This chapter addresses the sub-aim that is focused on developing a novel integrated web-based repository platform for CRC related omics data. This chapter provides a description of the developed resource and the features that CRC researchers can utilise to understand what is already known about the disease.
- Chapter 4: *Perturbation of protein-protein interaction network based on APC mutations in colorectal cancer*. The chapter discusses the prediction of genes which are essential for the proliferation of cancer cells when APC is mutated in CRC. In this chapter, I developed a novel network analysis method based on the node degree that integrates genomics and proteomics data to analyse the topological changes in a PPI network when APC is mutated and attempts to identify genes with the most topological changes as essential for the proliferation of cancer cells in CRC. Using this method, I identified new and already known genes which are essential for the proliferation of CRC cells. These genes were then validated using the Achilles dataset [12]. This chapter has been prepared as a manuscript and submitted for publication to a peer-reviewed journal.
- Chapter 5: *Integration of heterogeneous 'omics' data using semi-supervised network labelling to identify essential genes in colorectal cancer*. This chapter builds on the work done in Chapter 4 and I apply network theory-based methods to address the problem of high dimensionality in omics datasets and applyd network propagation to a semi-supervised machine learning technique to address the problem of heterogeneity in both omics datasets and cancer in identifying the essential genes in CRC. In this chapter, I identify known essential genes in CRC as well as a new set of genes that are likely to be essential in the study of CRC. This chapter has been submitted for publication and is currently under review in the Journal of *Computers and Electrical Engineering* special issue of "Recent Advances in Machine Learning and Artificial Paradigms".
- Chapter 6: *Physical coherence and network analysis to identify novel regulators of exosome biogenesis*. In this chapter, I apply physical coherence and network analysis to identify novel proteins that regulate the process of exosome biogenesis through the endosomal sorting complex required for transport (ESCRT) pathways

and help us further improve our understanding of exosome biogenesis in general. The chapter describes a network-based method that is applied in the analysis of the ESCRT machinery by eliminating the bias that exists in PPIs due to false positives stemming from experimental errors in techniques used to identify them and study biases. In this chapter, we identify STAMBP and NEDD4 as potential novel regulators of exosome biogenesis. This chapter is in preparation as a manuscript for submission to a peer-reviewed journal.

- Chapter 7: *General discussion.* This chapter provides an overall discussion of the findings of this chapter together with the implications of the findings on exosome biogenesis and cancer-related studies. Other issues discussed include the challenges faced as well as the future direction of the thesis.

Chapter 2

Background and related work

2.1. Introduction

This chapter highlights the use of networks in analysing omics data such as protein-protein interactions in cancer diseases. Several types of research have looked at the origin, progression and metastasis of cancer. Moreover, as tools for generating omics data have become cheaper over the years, vast quantities of heterogeneous data are today easily generated at a fraction of what it would have cost a few years ago. To study such vast quantities of data, researchers use networks and network theory in system biology to study the interplay of molecules in normal and cancerous conditions. Methods and tools have therefore been developed with varying degrees of success. The chapter evaluates some of the methods and tools and identifies some of the limitations associated with these tools and methods. The chapter further provides an overview of several studies related to the research aims which includes system biology, interactomes, exosomes and exosome biogenesis, networks and network theory, cancer - specifically CRC, and the computational analysis of PPI-related data to infer essential disease-associated genes.

2.2. Colorectal cancer

Colorectal cancer, also known as bowel cancer, is a form of cancer that originates in the colon or rectum section of the large intestine. It has the third highest number of incidences of all cancers in the world, and if not detected and treated early, CRC has one of the highest rates of cancer-related mortality in the world [13, 14]. The pathogenesis of CRC is still a subject of extensive research in the field of oncology, nonetheless, like in other cancers, the classic view is that alterations to the DNA essentially cause CRC resulting in the acquisition of a set of characteristics which lead to the abnormal and unregulated growth of cells. These characteristics were first put across by Hanahan and Weinberg [15], [16] and comprise the evasion of programmed cell death, self-sufficiency in growth signals, insensitivity to growth inhibitory signals, and sustained angiogenesis (ability to form blood vessels). Others are tissue invasion and metastasis (spreading of cancer to other tissues), limitless replicative potential, deregulating cellular energetics (modification of cellular metabolism) and avoiding immune destruction (cancer cells evading immunological destruction). Figure 2-1 provides a summary of the characteristics of cancer.

According to the World Cancer Research Fund International (WCRF), 95% of CRCs are adenocarcinomas while others are mucinous carcinomas and adenosquamous carcinomas. The pathogenesis of adenocarcinomas in CRC is preceded by the development of growths in the linings of the intestine called polyps due to the accumulation of either inherited or acquired somatic mutations which then transform glandular epithelium into adenocarcinomas [17-19]. Upon further accumulation of mutations, the adenocarcinomas become invasive and metastasise to other organs such as the liver. The majority (>70%) of CRCs are sporadic, and only about 20% of CRCs are hereditary [20]. Hereditary CRCs are due to rare, high risk, susceptibility syndromes such as Lynch syndrome (LS) and familial adenomatous polyposis (FAP) [20]. Sporadic CRCs, on the other hand, are due to the accumulation of genetic mutations in several genes. Three genetic mechanisms underlie predisposition to sporadic CRCs [20, 21] and consist of: chromosomal instability (CIN), microsatellite instability (MSI) and the CpG island methylator phenotype (CIMP) pathways [22]. Of the three, the CIN pathway is implicated in the majority of sporadic CRCs [22, 23].

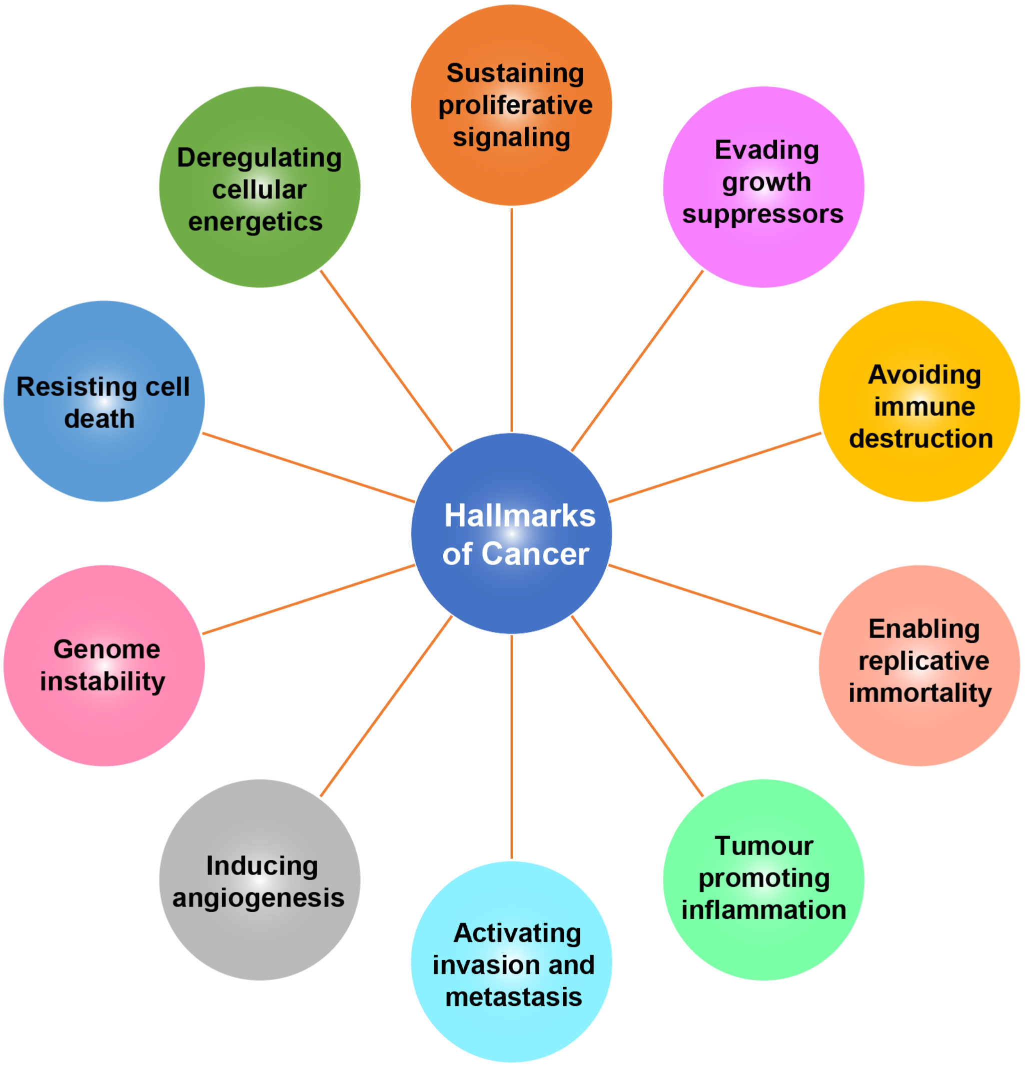

Sustaining proliferative signaling
Evading growth suppressors
Avoiding immune destruction
Enabling replicative immortality
Tumour promoting inflammation
Activating invasion and metastasis
Inducing angiogenesis
Genome instability
Resisting cell death
Deregulating cellular energetics
Hallmarks of Cancer

Figure 2-1: Hallmarks of Cancer. The hallmarks of cancer are the characteristics that cancerous tumours acquire, allowing for the unregulated growth of tumours and ultimately metastasis.

The classical model for the tumorigenesis of CRC was first proposed by Fearon and Vogelstein [24] and consists of a multistep accumulation of mutations in multiple genes. The first step comprises the accumulation of mutations in adenomatous polyposis coli (APC), a tumour suppressor gene which leads to the loss of functionality in APC [25]. The loss of APC functionality is coupled with the activation of mutations in KRAS, an oncogene as well as further mutations in PIK3CA, TP53 and transforming growth factor-β pathways [22, 24, 26]. Over the years, the model has been revised to include more genes (≈80) out of which 15 genes are considered key drivers of tumorigenesis in CRCs [22, 27]. Some of these include: APC, TP53, KRAS, PIK3CA, FBXW7, SMAD4, TCF7L2, NRAS, CTNNB1 (β-catenin), SMAD2, FAM123B, and SOX9 [20, 27-29]. While the model first proposed by Fearon and Vogelstein [24] has since undergone revisions, research has shown that >80% of sporadic CRCs have APC mutations [30]. Interestingly, it has also been shown in-vitro that tumorigenesis is only observed when APC mutations are present, even if other gene mutations such as those in KRAS are present [31].

There are several studies in the literature which document the role of APC in CRC [17, 18, 20, 21, 25, 30-34]. APC functions by regulating other genes such as CTNNB1 of the Wnt signalling pathway, a pathway that regulates cellular behaviours such as cell migration, cell polarity, and organogenesis. APC, therefore, indirectly regulates the Wnt signalling pathway through its regulation of CTNNB1, thereby regulating functions such as cell adhesion and migration, and signal transduction as well as other functions like microtubule assembly and chromosome segregation. Consequently, in CRC, mutations in APC lead to the loss of its functionality which, in turn, leads to the hyperactivation of the Wnt signalling pathway, the main characteristic of CRC [35, 36].

2.3. Protein-Protein Interactions (PPIs)

Life as we know it starts with a fertilised egg which then progresses into a collection of identical cells that gradually develop into a full-blown individual. This process of development and growth is a complicated process which is always taking place in our bodies and requires a sophisticated system of checks and balances. In addition, to ensure that the correct path for development is followed, cells communicate with each other and cells that are considered not to be necessary anymore should be removed with minimal disturbance to the other cells. Organs progressively develop their own blood supplies as well as mechanisms to repair any damage to these supplies. For this to be possible, there is a need for organs to communicate. All this is achieved by switching genes on and off in a synchronised manner as organs and systems develop. Once the development process is complete, the next step is the maintenance and repair of damage to the fully developed tissues.

The network of on and off switches is vital in the regulation of cellular behaviour; if there is a failure in these controls, it results in the uncontrolled growth of cells as occurs in cancers [2]. A gene is a fundamental unit around which DNA is organised. Proteins, on the other hand, are one of the expression of genes. Proteins are encoded for by genes and produced when a gene is transcribed onto messenger Ribose Nucleic Acid (mRNA).

Interactome networks in systems biology are the interactions between cellular components [37]. They give a global picture that is useful in understanding how interactions between molecules influence cellular behaviours [38]. Typical examples of interactomes are protein-protein interactions, virus-host network, transcriptional regulatory networks, metabolic network, and disease network. In a gene-regulatory network, nodes denote transcription factors which are depicted by circular nodes; and edges depict the physical binding between the two [39]. Disease networks are networks which depict the link between disorders and genes that are known to be associated with the diseases [40]. Diseases in disease networks are denoted by nodes and edges denote gene mutations linked to the disease [8]. A virus-host network, on the other hand, is modelled by viral proteins depicted as square nodes or host proteins as round nodes while edges depict physical interactions between the two. In a metabolic network, nodes depict enzymes, and edges depict metabolites that are products or substrates of the enzymes. The scope of this thesis revolves around PPIs.

Proteins are macromolecular structures that make up the working machinery of the cellular system. They handle a range of functions within an organism, such as acting as molecular motors, catalysing reactions, transportation, traversing of membranes producing regulated channels, transmission of DNA information to RNA, and signalling. However, proteins do not work independently but interact with other proteins, DNA, RNA, and other small molecules within cells. A PPI is the result of two or more proteins binding together purposely to carry out a specific biological function in a cell [41]. PPI bonds are formed by a combination of hydrophobic bonding, van der Waals forces, and salt bridges at specific binding domains on each protein. The PPIs form complexes which then conduct many of the molecular processes in the cell, such as DNA replication, metabolic signalling, gene-regulation and immunity. PPI networks offer an overall depiction of cellular function and biological processes within an organism. They offer an essential network which is critical for the flow of vital information for bio-molecular functions and overall cellular processes [42].

The complexity of PPIs is simplified through the representation of PPIs as graphs which are composed of nodes as proteins and edges symbolise physical, biochemical and functional interactions between the two proteins [37] as shown in Figure 2-2 . Similarly, biological networks such as PPI networks can be likened to communication networks in that they both have the properties of being scale-free and having a ‘small-world’ [43]. PPI networks, therefore, can be used to demonstrate the evolutionary aspects of proteins [44], to improve protein function annotation [45] and to represent the modular organisation of a cell [46].

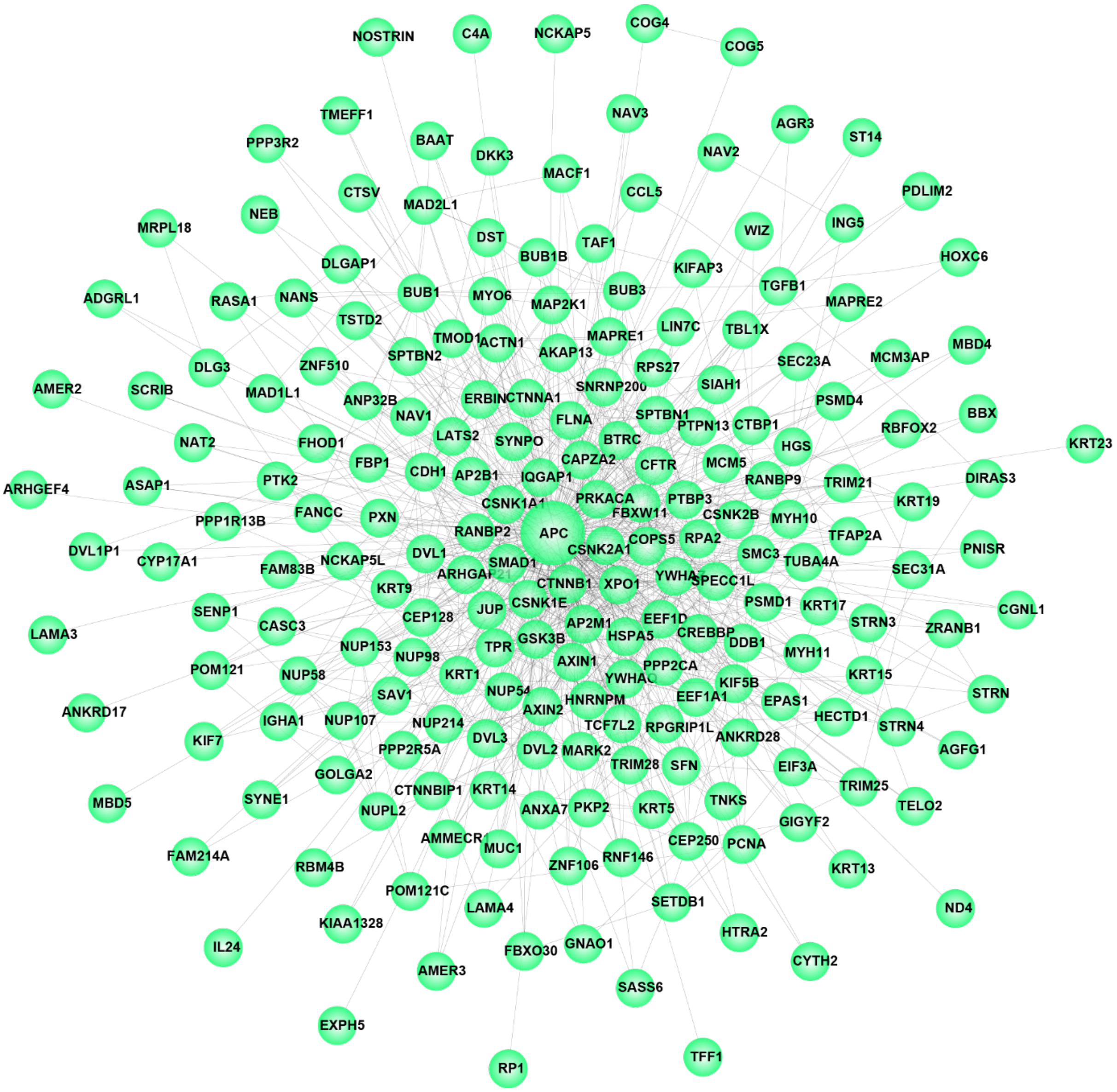


Figure 2-2: Example of APC PPI Network. The interaction shows APC and its interacting partners

Types of PPIs

The levels or types of PPIs are classified differently depending on their biological features. In a review paper by De Las Rivas and Fontanillo [47], three levels are highlighted: co-interacting proteins, correlated proteins, and co-located proteins.

Co-interacting proteins are considered to be the physical interactions; they are further categorised as stable (permanent) or transient interactions, with both types being either strong or weak. Stable interactions constitute protein complexes that carry out either a structural or functional biomolecular role. These proteins make up the subunits of the complexes. Examples consist of nuclear pole subunits and subunits of ATPase. Transient interactions, on the other hand, are considered to control most of the cellular processes, are temporary in nature and often need specific conditions to promote interactions, such as conformational changes and phosphorylation. They come together when certain conditions are met to carry out a biomolecular function. Examples include most of the proteins involved in signal transduction.

Correlated proteins reflect those proteins that are involved in the same biomolecular activities but do not physically interact. These can be metabolically correlated or genetically correlated. Metabolically correlated proteins are those proteins found in the same metabolic pathway. Examples of such are mostly *enzymes* implicated in Krebs cycle. Genetically correlated proteins are those that are encoded by genes that are co-expressed or co-regulated. Examples include those proteins that regulate a portion of the cell cycle.

Co-located proteins refer to those that are localised in the same organelle. These can be located in the same cellular space such as those proteins in lysosome, or they can be found on the same cellular membrane such as those that act as receptors in the plasma membrane.

From the above, two proteins can interact under one of the types of association. The defined interactions are not exclusive; two proteins can interact using either of the associations at any given time.

Mapping of PPI networks

Developing a network of all probable physical PPIs or the ‘interactome’ is a significant part of systems biology. Proteome-scale interactome network mapping can be traced back to the mid-1990s through research on organisms such as *Escherichia coli*, *Drosophila melanogaster, and Saccharomyces cerevisiae* [48-51]. Three main approaches can be used to map interactome networks [37, 52].

The first approach involves the curation of existing data from published literature often obtained for one or just a few types of physical or biochemical interactions [53]. The increasing and wide availability of published scientific papers compounded by the addition of detailed genetic information from the human genome sequencing project has resulted in an enormous knowledge-base all contained within the scientific literature. This approach involves the development of text mining methods that can scan through this enormous repository of free-form, unstructured data (written articles) transforming it into highly structured information that can be used to deduce inter-relationships for proteins, diseases, or species. Relevant literature is first identified which is then followed by the extraction and classification of related terms or entities such as proteins, genes, diseases or pathways. The technique, therefore, reduces the complexity and ambiguity of large repositories of unstructured literature by identifying and creating relationships. While manual curation is a possibility, the massive amount of data that is available in the several databases can be a nightmare to a researcher to efficiently and effectively decipher relationships, for instance, PubMed alone as of 2014 contained more than 23 million citations from Medline, life science journals, and online books. Text mining (computer processing of large text) using computer algorithms takes the information overload off a researcher. Reviews by Cohen and Hersh [54] and Erhardt, et al. [55] identify several groups of biomedical text mining approaches as well as the pros and cons of each approach.

The second approach involves the use of computational methods in the prediction of interactome networks based on the structural, genomic and biological context of genes and proteins in completely sequenced genomes [56, 57]. The computational prediction of PPIs can be summed up in a two-step process: the first step is the mapping of PPIs and the second step deals with the comprehension of the mechanism by which the proteins interact and isolating the residues of proteins that are involved in the interactions.

The third approach and one of the earliest to be used in system biology involves the use of experimental techniques to identify PPIs. The experimental techniques are divided into low-throughput screens and high-throughput screens [58]. Low-throughput (LTP) experiments are the yardstick for measuring interactions due to their reliability. Examples of low-throughput techniques include: affinity precipitation, dosage lethality, biochemical assays, affinity chromatography, synthetic lethality and structure [59-61]. The curation of interactions detected in LTP screens is done by manually examining the publications, making the detection of PPIs using LTP extremely difficult. The second sets of experimental approaches are the high-throughput (HTP) experiments. Examples of HTP techniques include: yeast-2-hybrid (Y2H) [56, 61, 62], mass spectrometry-based methods [63-68], protein chips (microarrays) [69] and LUMIER assays [70].

Each of the methods discussed has varying degrees of effectiveness and reliability in detecting PPIs. Cusick, et al. [71] and Turinsky, et al. [72] argue that while literature curated maps can be easily curated from the available literature, they have the disadvantage of variability in quality, a lack of both systematisation and published data. An analysis of computational-based approaches for generating networks by Plewczyński and Ginalski [73] concluded that while computational-based methods are faster and more efficient as well as having the ability to generate large numbers of nodes and edges, computational-based methods tend to be defective because of their reliance on secondary information. High-throughput maps, on the other hand, are difficult and expensive to conduct but tend to produce highly reliable and comprehensible network maps. According to Hosur [58], experimental techniques, on the other hand, suffer from some limitations such as high false positives and negatives as well as low sensitivities when compared to computational approaches. Figure 2-3 provides a summary of the various methods that are used to map PPIs [74].

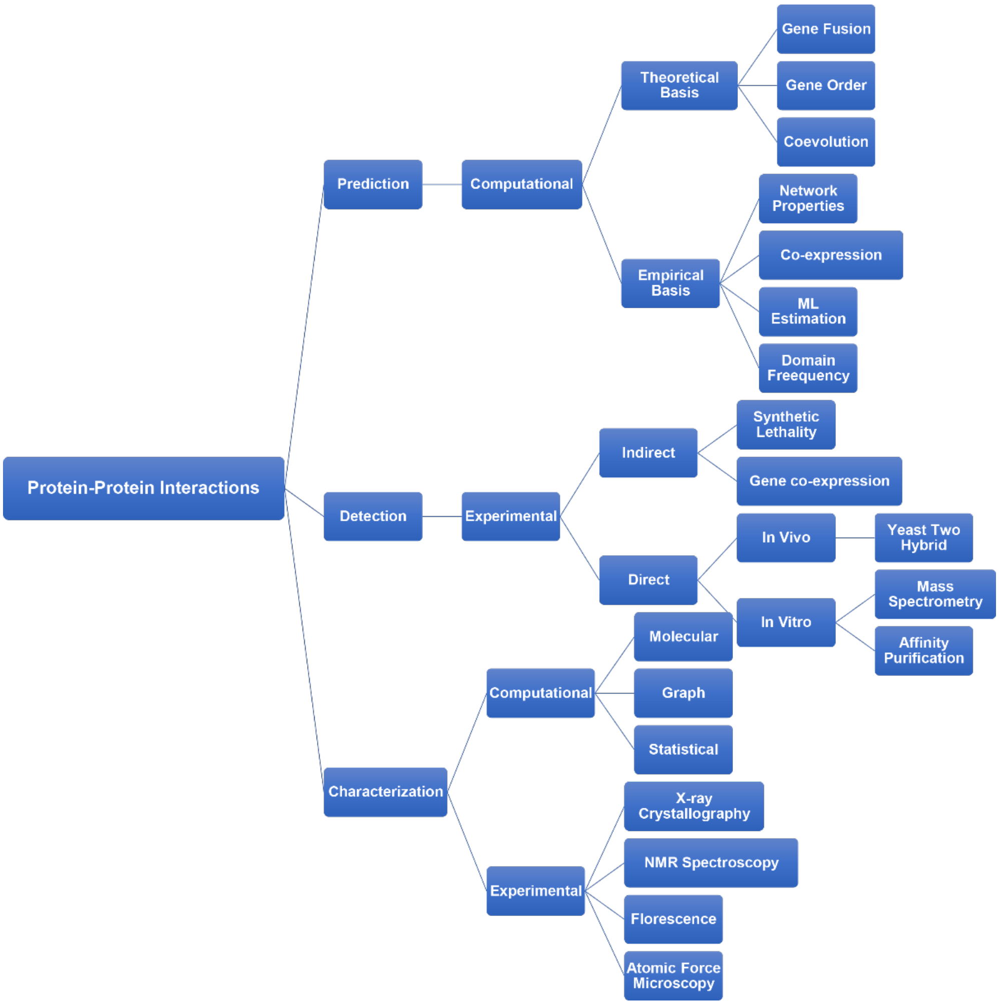


Figure 2-3: PPI mapping methods. The different types of methods used to detect interactions between proteins can be categorised into prediction, detection and characterisation.

2.3.1. PPIs in Disease

One of the challenges faced by researchers is how to understand the molecular mechanisms that precede a disease. Because of the central role of proteins in the cellular system, protein interactions have been found to play a regulating role in mechanisms that lead to both physiological and pathological states in organisms [74]. Some diseases are born from changes that affect the binding interface or lead to a biochemical dysfunction causing allosteric changes in proteins.

Our knowledge about the disease is usually associated with us knowing the genetic basis of diseases. With the introduction of Mendelian genetics in the 1900s, there have been efforts to isolate genes linked to diseases such as cystic fibrosis, Huntington disease and breast cancer susceptibility using the process of gene cloning whereby a gene is isolated based on its position on the chromosome [75]. However, it has been shown that even with Mendelian diseases, there is no direct or clear correlation between the mutations and the resulting disease symptoms. This observation is alluded to by such factors as pleiotropy (when a single gene encodes multiple phenotypes), the regulation of a gene by other genes, and environmental factors (such as diet, infection by bacteria, exposure to chemicals). All these factors, therefore, make it difficult to associate gene mutation to phenotypic expressions exactly.

In addition to the genetic basis of a disease, knowing the molecules and molecular mechanisms that trigger and regulate the perturbed biological process is important to understand the pathogenesis and progression of disease [74]. Nonetheless, trying to deduce the molecular mechanisms that lead to a disease is an even greater task than inferring the genetic basis of a disease. Networks using PPIs provide us with the opportunity to infer the molecular mechanisms behind the diseased states of an organism since they are involved in several cellular processes. PPIs act as a key source of molecular information since their interactions are involved in a wide range of activities such as signalling, immune response, metabolic and gene-regulatory networks [74].

Studies by Gonzalez and Kann [74] as well as those by Ideker and Sharan [76] proposed the use of PPIs as potent key targets for studying the molecular basis of diseased biological states. Diseased states have the potential to change interactions between proteins and their interacting partners such as DNA and ultimately lead to the formation of new undesired interactions, protein misfolds or the enabling of pathogen-host protein interactions. By studying these interactions, we can, therefore, find novel pathways involved in diseases.

Furthermore, PPI subnetworks tend to cluster together proteins that interact in functional complexes and pathways [6]. Studies by Hallock and Thomas [77] and Ideker, et al. [78] have shown that the pathways found from PPI networks can be used to generate hypotheses that can be used to study diseases. PPIs are therefore an opportunity to gain knowledge into disease state pathways and molecules. By applying PPIs and pathway analysis, researchers have been able to infer several features that can be associated with the disease. For example, studies by Goh, et al. [40] found that by studying the human disease network, it was shown that genes associated with the same disease or disorders had a higher likelihood that their products would have a physical interaction. Goh, et al. [40] further showed that genes which are essential tend to code for hub proteins and are widely expressed in most tissues thereby showing the importance of disease genes in the human interactome. On the contrary, it was shown that not all disease genes are essential and there is usually no indication of them ever encoding for hub proteins; their expression patterns also showed that they are found in the functional periphery of the network. On the contrary, for diseases such as cancer, it has been found that disease genes tend to encode for hub proteins that are highly interconnected [44, 79]. This view is supported by Ideker and Sharan [76] who concluded that genes associated with a phenotype, function or progression of disease are not randomly positioned in a network but tend to exhibit higher connectivity, cluster together and are located in the central network location. Lim, et al. [80] also showed that diseases which are aetiologically different tend to show similar symptoms since different biological processes share the same pathways.

The study of PPIs is, therefore, an important element in our quest to understand cancer. Gonzalez and Kann [74] summarised the five ways in which PPIs can be applied to the study of disease. Firstly, PPIs can be used to differentiate healthy from diseased states by developing interaction networks in varying conditions. For instance, Charlesworth, et al. [81] applied this principle in their study to identify the alterations in the canonical pathway and interaction networks when humans are exposed to cigarette smoke. Li, et al. [82] on the other hand, developed a computational method that can predict CRC-related genes by integrating gene expression profiles and the shortest path analysis method to PPIs. Their results showed that by using PPIs, they identified more cancer-related genes than they did by computing the differential gene expression between normal and diseased samples.

Secondly, PPIs can be used to predict genotype-phenotype relationships which, in the process, can help in the inference of novel disease genes. Gene-disease association studies commonly study interacting disease-associated proteins to identify disease-causing genes. For example, in a study by Gandhi, et al. [52], it was found that an interacting disease-associated protein was encoded by mutated genes in inherited genetic disorders.

Thirdly, the genes associated with interacting proteins can be used to study mutations that take place leading to alterations in the interactome in healthy individuals. These mutations can also be used to develop new interactions that appear in diseased states. Rossin, et al. [83] applied a genome-wide association study (GWAS) and developed a PPI network for genes within a given loci from which they established that there were significant interactions between protein products of associated genes.

Fourthly, PPI networks can be used to establish pathways as well as disease-subnetworks that are activated because of disease and then used as new biomarkers for identifying diseases. Chuang, et al. [6] identified a set of sub-network biomarkers that distinguish metastatic tumours from non-metastatic tumours in breast cancer patients. Fifthly, by identifying key nodes in disease networks established from PPI networks, drugs can be designed specifically to target nodes of interest.

2.3.2. PPIs in cancer

In cancer, PPIs form the signalling pathways needed for the transmission of pathophysiological signals which, in turn, lead to tumorigenesis, tumour progression, invasion, and metastasis [84]. Combinations of genetic and epigenetic (genetic changes not caused by changes to DNA but are due to external changes) alterations determine the potential of cells becoming cancerous. This is achieved through a sequence of signal networks with PPIs acting as the basic units forming these networks. PPIs play vital roles in the initiation of cancer by connecting networks that transmit oncogenic signals as well as conducting other roles in driving and sustaining the growth of cancer cells. The role played by this cascade of networks is summarised by Ivanov, et al. [84] in Figure 2-4. In cancer progression, cells acquire the ability to evade growth suppressors. PPIs again play vital roles in neutralising tumour suppression mechanisms whereby the tumour suppression mechanisms are hijacked by viral oncoproteins which induce tumours [15]. Furthermore, PPIs also facilitate the acquisition of other hallmarks of cancer as shown in Figure 2-4. Therefore, mutations leading to cancer result in the reprogramming or alteration of PPI networks leading to the formation of PPIs that facilitate distinct features of cancer or play key roles in other multiple characteristics of cancer.

Against this background, probing the interface properties of cancer-related proteins is valuable to the understanding of biological processes and protein functions that underlie cancer. With the help of high-throughput experimental data, techniques have been established that are used to identify PPIs [44]. Nonetheless, the inference of genes associated with a disease requires the analysis of thousands of genes across a cohort of potential candidates. This process, however, requires the use of efficient methods. Several computational methods have therefore been developed to aid in linkage analysis and gene-disease association studies. The majority of these methods are based on functional and sequential differences between disease-causing and non-causing proteins. Other methods combine several sources of data such as gene ontology (GO) annotation, gene expression and disease phenotype representation from various databases [43, 74].

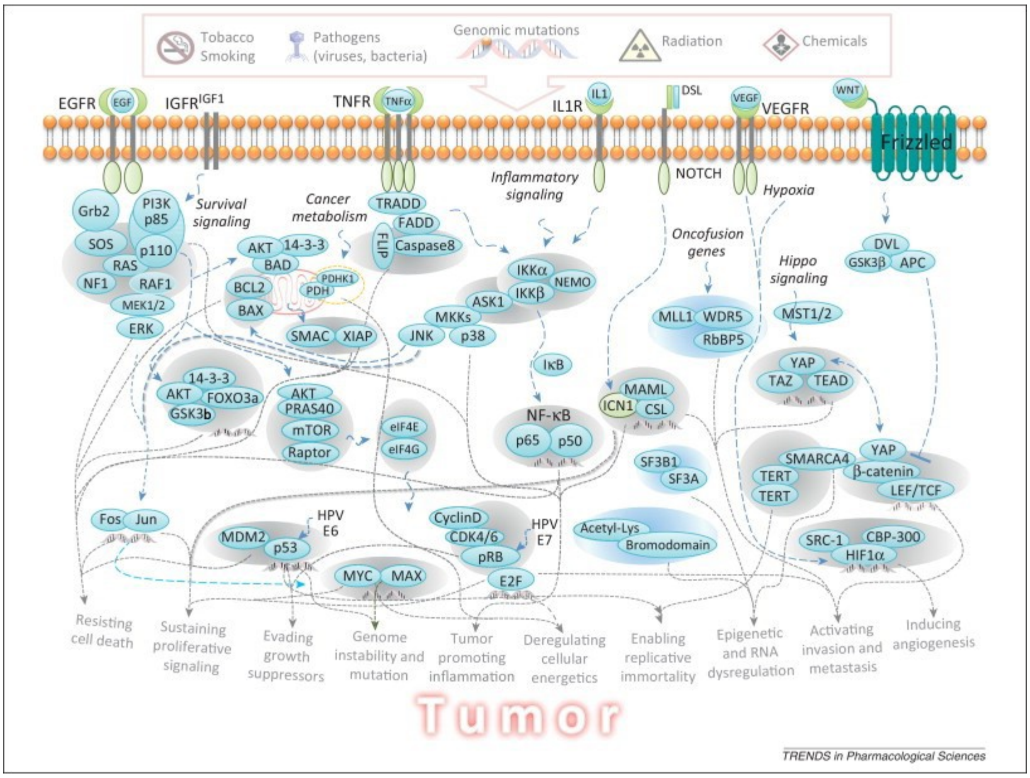


Figure 2-4: PPIs in Cancer. Examples of PPI networks that drive the acquisition and development of cancer hallmarks by Ivanov, et al. [84]

2.4. Bioinformatics tools and methods in cancer studies

The completion of the human genome project in 2003 together with the development of new high-throughput technologies such as mass spectrometry (MS), next generation sequencing (NGS) (RNA-seq) and single-cell RNA sequencing (scRNA-seq) [85] have led to the generation and accumulation of a rich quantity of data. To contend with the analysis and interpretation requirements of these data, bioinformatics, an interdisciplinary field of science that combines the fields of computer science, statistics, and biology to the development of computational tools has become an essential element of science. Bioinformatics tools and methods are used to capture, analyse, and interpret biological data. In cancer-related studies, bioinformatics continues to play a key role in whole genome analysis, drug discovery and personalised medicine, biomarker prediction and system biology in general. This section discusses some of the existing tools and methods that have been applied in understanding the physiological mechanisms of PPIs in diseases.

2.4.1. Network tools for the analysis of proteomic data

This section is presented as a manuscript which was published as a book chapter in *Proteome Bioinformatics* part of the *Methods in Molecular Biology* book series (MIMB, volume 1549) (Chisanga, et al. [86]).

This section gives an overview of the application of network theory together with protein-protein interactions in analysing proteomic data, a key component in understanding the physiological mechanism of disease. Existing tools and resources for the capture, storage and analysis of proteomic-related data are also discussed.

Network tools for the analysis of proteomic data

David Chisanga[1], Shivakumar Keerthikumar[2], Suresh Mathivanan[2] and Naveen Chilamkurti[1,*]

[1]Department of Computer Science and Information Technology, La Trobe University, Bundoora, Victoria, 3086, Australia

[2]Department of Biochemistry and Genetics, La Trobe Institute for Molecular Science, La Trobe University, Melbourne, Victoria, 3086, Australia

*To whom correspondence should be addressed

Dr. Naveen Chilamkurti

Department of Computer Science and Information Technology,

La Trobe University,

Bundoora, Victoria 3086, Australia

Tel: + 61 03 9479 1269

Fax: +61 03 9479 3060

Email: N.Chilamkurti@latrobe.edu.au

Abstract

Recent advancements in high-throughput technologies such as mass spectrometry have led to an increase in the rate at which data is generated and accumulated. As a result, standard statistical methods no longer suffice as a way of analysing such gigantic amounts of data. Network analysis, the evaluation of how nodes relate to one another, has over the years become an integral tool for analysing high throughput proteomic data as they provide a structure that helps reduce the complexity of the underlying data.

Computational tools, including pathway databases and network building tools, have therefore been developed to store, analyse, interpret and learn from proteomics data. These tools enable the visualization of proteins as networks of signalling, regulatory and biochemical interactions. In this chapter, we provide an overview of networks and network theory fundamentals for the analysis of proteomics data. We further provide an overview of interaction databases and network tools which are frequently used for analysing proteomics data.

Keywords: Proteomics, Network theory, Protein-protein interactions, Network tools, Network analysis, Bioinformatics.

2.4.1.1. Introduction

In recent years, the development of high-throughput technologies such as next generation sequencing techniques in the field of genomics and tandem mass spectrometry in the field of proteomics and metabolomics has led to the birth of the omics study [87]. These techniques and tools involved in the study of functional genomics and other omics data have constantly helped in our understanding of cellular biology and have drastically reduced the cost of conducting omics related studies. The speed with which data are generated and disseminated today means that researchers can gain insight for the fraction of the cost when compared to past years. For instance, by using tandem mass spectrometry, two groups [88, 89] have developed the first draft of the human proteome.

However, with terabytes of proteomic data pouring into research centres every day, standard statistical methods for analysing data are becoming ineffective. Researchers are faced with the formidable task of how to take advantage of this heterogeneous data to gain insight in areas such as disease and drug development as well as answering questions such as; how can they characterise and manipulate complex interactome of basic elements such as genes and proteins? How can they visualise these interactomes and infer meaningful information from them?

Network theory has long played a fundamental role in disciplines ranging from computer science, sociology, engineering, and physics, to molecular and population biology [90]. In biology and medicine, network analysis methods are applied in areas such as drug target identification, prediction of a gene or protein function, protein complex or module detection, prediction of novel interactions and functional associations, identification of disease subnetworks, disease biomarker identification, and mapping of disease pathways [10]. Networks have long been used in a variety of fields to reduce the complexity of data [91, 92]. Computational tools, including pathway databases and network building tools, have been developed to store, analyse, and interpret biological networks [93].

This chapter provides an overview of the application of network theory in analysing and visualization of proteomic data by discussing various tools used for storage, analysis and interpretation of proteomic data through the use of biological networks with an emphasis on protein-protein interaction networks. To get started, we provide a brief background to both proteomics and network theory.

2.4.1.1.1. Background to proteomics

Coined by Marc Wilkins and colleagues [94] in the mid-1990s to mimic the terms "genomics" and "genome" respectively, proteomics is in essence a systems science whose aim is to identify and record the functions as well as structures of proteins in organisms. Proteomics is a systems science which involves not only the measurement of proteins but also the measurement of their expressions in a cell and the interplay of proteins, protein complexes, signalling pathways, and network modules.

Proteins are termed as the workhorses of cellular systems, they perform an array of cellular functions ranging from catalysing reactions, cellular transportation, transcription of DNA information to RNA and acting as molecular motors to signalling [95]. They perform these functions not on their own, but within large complexes where they interact with other molecules like proteins, DNA, RNA as well as with other small molecules. Because of their importance, a malfunction in key proteins can lead to serious pathological outcomes like cancer, metabolic imbalances, and neurodegenerative diseases. With significant ongoing research into protein functionality and their interactions with other molecules in understanding disease, research has turned to network theory concepts to model and study these interactions.

2.4.1.1.2. Background to network theory concepts

A network or a graph (in mathematics) is a collection of objects connected by lines. The objects are called nodes or vertices while the connections between the objects are called edges or links and are drawn as lines between points as shown in Figure 2-5

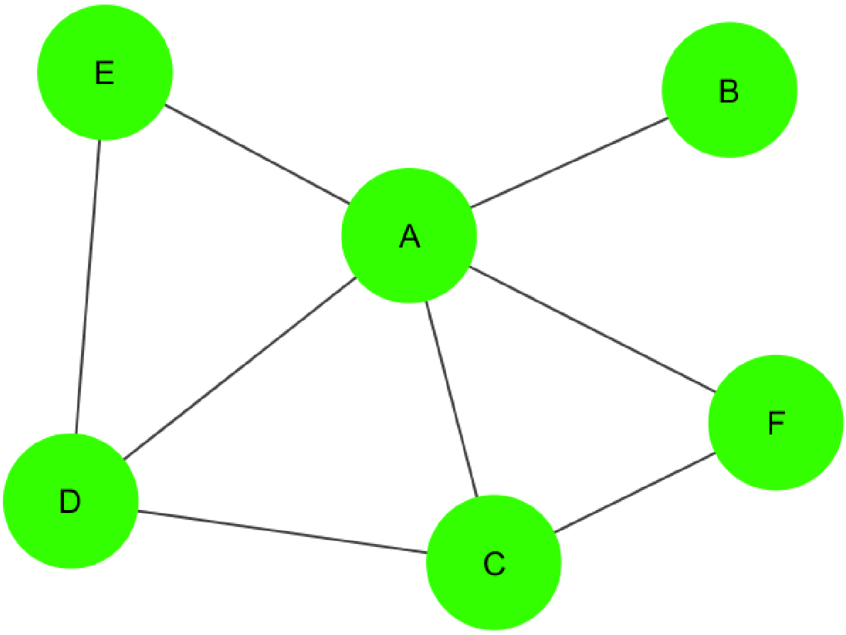


Figure 2-5: Example of an undirected network graph in which each node is connected by an edge that does not show the origin and destination by way of an arrow.

Formally, a network is a graph G defined as an ordered pair G= (V, E) where V is a set of nodes and E is a set of edges [90]. Nodes are said to be adjacent if they are joined by an edge while node ‘A’ is said to be a neighbour to node ‘B’ if adjacent to node ‘B’ and vice-versa. Edges between nodes can be undirected (Figure 2-5) or directed (Figure 2-6), as such a graph G= (V, E) is called undirected if an edge vv’ (where v and v’ are nodes in set V) in set E of edges implies that it is the same as edge v’v also in E; otherwise G is called directed. A directed acyclic graph, on the other hand, is a directed graph that contains no cycles. Finally, a graph is said to be connected if there is a path from any node to any other node.

Using the above network/graph concepts, researchers have used networks to reduce the complexity of systems thereby making it easier to draw conclusions from them. Networks are applied in various fields such as computer networks, social networks, and interactome networks in molecular biology research.

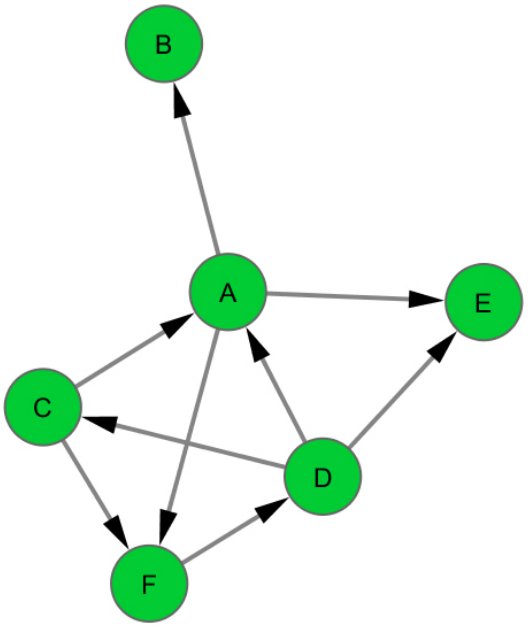


Figure 2-6: Example of a directed graph in which each node is connected by an edge with an arrow indicative of the direction of the relationship.

Interactome networks provide a global picture that is useful in understanding how interactions between molecules influence cellular behaviour [38]. It has been established that biological behaviour arises from the complex interactions between the cell's numerous molecules such as proteins, DNA, RNA and other small molecules. Common examples of interactomes in molecular biology are; protein-protein interactions, virus-host networks, transcriptional regulatory networks, metabolic networks, and disease networks. Protein-Protein Interactions (PPIs) form the backbone of signalling pathways, metabolic pathways and cellular processes required for normal functioning of cells [96].

The steps to perform proteomic analysis can be summed up by use of a flowchart as shown in Figure 2-7, it involves identifying a set of target proteins of biological interest needs to be studied and then followed by retrieval or identification of interacting partners from various interaction resources discussed below. An interaction network is then generated and integrated with any existing knowledge such as gene ontology (GO) enrichment, biological pathways or differential gene or protein expression. A topological analysis of the network is then performed using metrics such as degree, degree centrality or betweenness centrality which is further followed on by downstream analysis to identify network variations, functional enrichment of identified modules or tissue specificity.

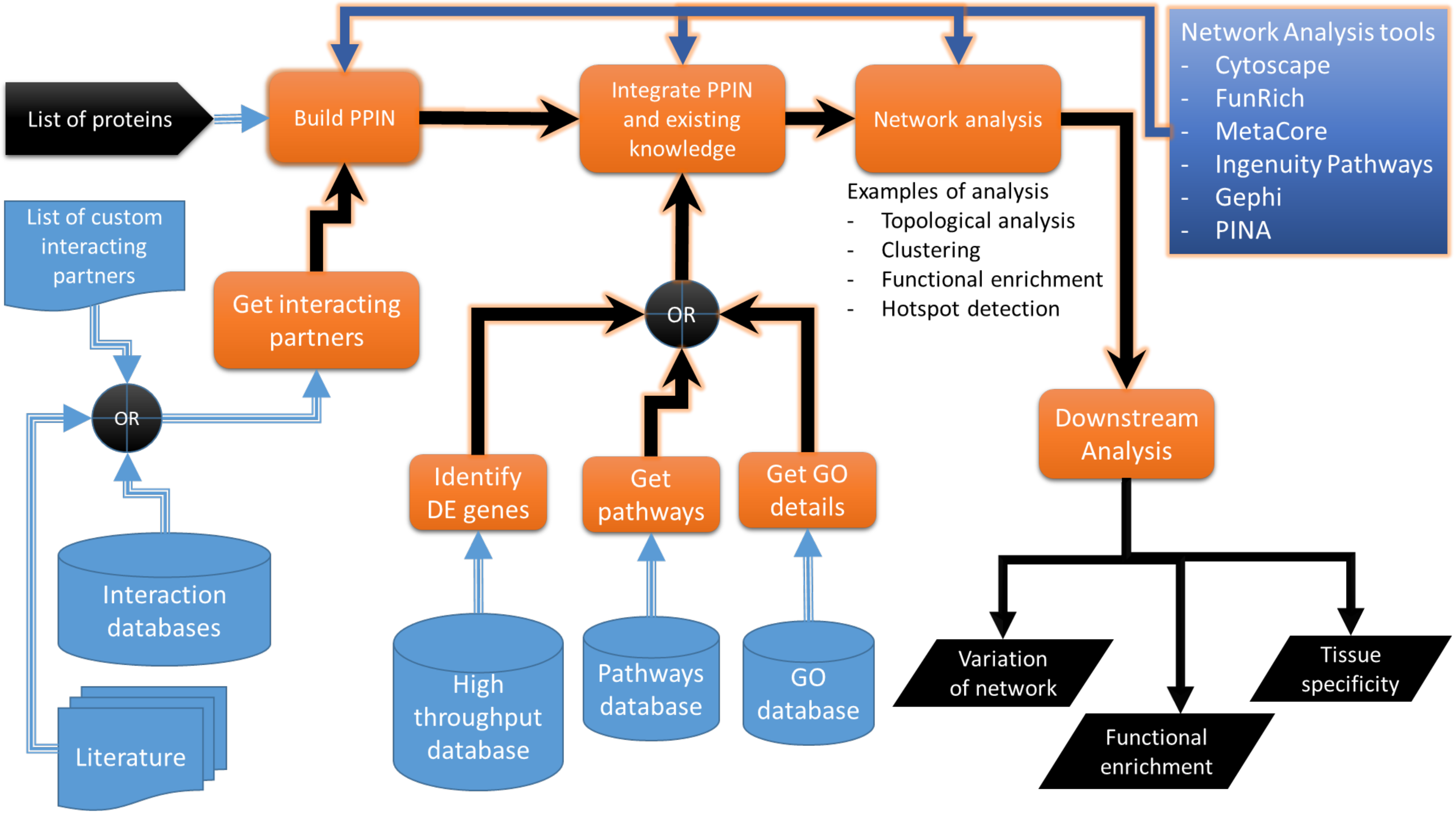


Figure 2-7: Summary representation of the steps involved in analysing proteomic data using network theory concepts. The data types required and from where they can be sourced are also shown. An example of the expected outputs from the network analysis are also shown.

2.4.1.2. Protein-Protein Interaction databases

The mappings of proteins and their interacting partners have been curated by various groups and deposited into online databases. These databases are typically web-based resources that serve as archives of information pertaining to the mapping of protein interactions, functional enrichment (GO enrichment) and pathway details. These databases act as sources of protein mapping information in network analysis. The most widely used PPI databases include; Human Protein Reference Database (HPRD) [97], Molecular Interaction Database (MINT) [98], Biological General Repository for Interaction Database (BioGRID) [99], Search Tool for Recurring Instances of Neighbouring Genes/Proteins (STRING) [100], Database of Interacting Proteins (DIP) [101], Biomolecular Interaction Network Database (BIND) [102] and the IntAct molecular interaction database (IntAct). Depending on the database, the annotations may be based on experimental observations while other databases such as STRING can have a high proportion of predicted and literature mined interactions. Below, we briefly discuss the most commonly used databases while Table 2-1 provides a summary of these and other database resources with protein-protein interaction mappings.

2.4.1.2.1. BioGRID

The Biological General Repository for Interaction Datasets (BioGRID) is an open, accessible web-based repository of genetic and protein interaction mappings which have been curated from the primary biomedical literature of humans and other major model organism species [99]. As of May 2016, the database housed over one million (1,000,000) protein and genetic interactions curated from over fifty-six thousand (56,000) high-throughput datasets and individually focused publications for major model organisms.

BioGRID features an easy to use web interface with a search tool which users can use to search against the database, the search results then show the interacting partners, interactor details and a graphical network visualisation of the interacting partners. Users can then manipulate the network by either changing the network layout or filtering through the network by node degrees. In addition, users can also download custom defined or entire interaction datasets for offline network analysis and downstream analysis. BioGRID also features online tools and resources that allow for the use of BioGRID data. A number of visualisation tools such as Osprey, Cytoscape, and GeneMania, data management tools like ProHits, plugins like BioGRID Tab File Loader Plugin for Cytoscape and BiogridPlugin2 for Cytoscape as well as web services BioGRID REST Service and PSICQUIC provide users with access to or can be used to analyse BioGRID data.

Table 2-1: Summary of database resources that house protein-protein interactions and their respective features

Resource	**Description**	**URL link**	**Reference**	**No. Proteins**	**No. Interactions**	**No. Organisms**
BIND	Biomolecular Interaction Network Database	http://bond.unleashedinformatics.com/	[102]	23,643	43,050	80
BioGRID	Biological General Repository for Interaction Datasets	http://thebiogrid.org/	[99]	56,105	553,827	175
HPRD	Human Protein Reference Database	http://www.hprd.org	[97]	30,047	41,327	1
IntAct	IntAct Molecular Interaction Database	http://www.ebi.ac.uk/intact/	[103]	89,716	356,806	131
MINT	Molecular INTeraction database	http://mint.bio.uniroma2.it/mint	[98]	35,553	241,458	144

STRING	Search Tool for the Retrieval of Interacting Genes/Proteins	http://string-db.org/	[100]	9,643,763		2,031

2.4.1.2.2. Human Protein Reference Database

Human Protein Reference Database is a web-based resource that houses experimentally derived human proteome information [97]. It is one of the most comprehensive collection of human proteome information resource available online. It houses information pertaining to; protein-protein interactions, post-translational modifications and tissue expression. As of May 2016, the database housed over thirty thousand (30,000) protein entries, over forty-one thousand (41,000) protein-protein interactions, ninety-three thousand (93,000) Post-Translational Modifications (PTMs), one hundred and twelve thousand (112,000) protein expressions, twenty-two thousand (22,000) subcellular localisation details, four hundred (400) domains and with over four hundred and fifty-three thousand (453,000) PubMed links to publications.

The landing page of HPRD provides a range of features ranging from a querying functionality, BLAST feature to a browse feature. Users can query the database using the query page through a number of search options, the results are then displayed using graphical visual displays and are categorised into protein information, PTMs, protein length, and protein-protein interactions. Users can similarly get protein information through the browse page where the information is grouped into molecular classes, domains, motifs, PTMs and based on localisation. HPRD further includes a Basic Alignment Search Tool (BLAST) which allows users to search against the database based on the provided protein or nucleotide sequence. Other features included are a phosphor motif finder tool which searches across user submitted protein sequence for the presence of over 300 phosphorylation-based motifs listed in HPRD. HPRD also provides tab delimited files for binary protein-protein interactions which users can download for offline processing and further download stream analysis.

2.4.1.2.3. Molecular INTeraction database (MINT)

The Molecular INTeraction database [104] is a web-based resource that stores physical interactions between proteins of model organisms that have been curated from the scientific literature. As of May 2016, MINT had over two hundred and forty-one thousand protein-protein interactions (241,000), thirty-five thousand (35,000) proteins and over five thousand PubMed links to publications.

MINT data can be downloaded in several formats such as PSI-ML, tab-delimited and MINT flat file formats. Otherwise, users can use the search feature that allows users to search the MINT database. Users can search the database using several options such as by gene name, protein accession number or any 6-characters keyword. A user defined list of proteins can furthermore be uploaded and used to generate a network visualisation based on the information in the database.

2.4.1.2.4. Biomolecular Interaction Network Database

The Biomolecular Interaction Network Database [102] is a web-based resource for PPI data and was one of the earliest resources for biomolecular interactions (proteins, genes etc.), molecular complexes and pathways. BIND initiated by the University of Toronto as part of the Biomolecular Object Network Databank (BOND) has since been acquired by Thomson Reuters. BIND provides tools for data specification plus a database which is accompanied by data mining and visualization tools.

2.4.1.2.5. IntAct molecular interaction database

IntAct [103] is an open-source web-based molecular interaction database that catalogues data curated from the scientific literature or from direct data depositions. As of May 2016, IntAct had over five hundred and ninety-one thousand molecular interactions, and ninety-one thousand interactors sourced from over fourteen thousand publications.

Using IntAct users can explore the fine details of the mechanism by which a specific protein binds to protein partners or use the entire interactome of an organism to perform a network analysis of large-scale omics experiment. The front-end of IntAct features a search tool that can be used to search against the IntAct database. Users can then view the interacting partners, interaction details and a graphical presentation of the network.

2.4.1.2.6. Search Tool for Recurring Instances of Neighbouring Genes/Proteins (STRING)

STRING is a freely available web-based biological database that houses information on experimentally derived and predicted protein-protein interactions for a number of organisms. This information has been curated from various sources, including experimental data, computational prediction methods, and published literature. STRING holds over one hundred and eighty-four (184) million interactions, nine million (9,000,000) proteins from over two (2,000) thousand organisms.

STRING provides an easy to use web interface that allows users to quickly search for a protein of interest, visualize and download interaction data. It further has a Cytoscape plugin which allows users to directly access the STRING database from Cytoscape. The interaction data returned from STRING is weighted and allows for the calculation of confidence scores for each interaction. In addition, STRING has capabilities that allow it to connect to other databases and consequently perform literature mining. It also includes a capability that allows for the drawing of simple protein networks based on the provided list of genes and the available interactions in the database.

2.4.1.3. PPI data exchange formats

Interaction networks are represented in a number of different file formats, the most widely used formats are; tab delimited text (.tab or .txt format), excel workbooks (.xls format), simple interaction file (SIF or .sif format), nested network format (NNF or .nff format), graph markup language (GML or .gml format), XGMML (extensible graph markup and modelling language), SBML, BioPAX, PSI-MI level 1 and 2.5 formats. All the interaction repositories provide at least one of these formats as a way to download interaction data.

2.4.1.3.1. Delimited text and excel workbooks

The delimited text and excel workbook file formats are the most basic and widely used for working with interactive data and are supported by most if not all network analysis tools. Tables in these files can contain network and edge (interaction) attributes or values such as the confidence of an interaction. With these types of files, users can specify the columns for source and target nodes as well as interaction types, and edge attributes when importing network data into an analysis tool.

2.4.1.3.2. Simple Interaction Format (SIF)

This format allows for the construction of a network from a list of interactions by easily merging different interaction sets into a larger network. Each line in a SIF file specifies a source node, a relationship (or edge type), and, one or more target nodes as shown in the following example.

nodeA <relationship type> nodeB

nodeC<relationship type>nodeB

2.4.1.3.3. Nested Network Format

This format is simple and similar to the SIF format except it allows the option of nesting a network into a single a node. An interaction is specified by either of two possible formats [105, 106];

- A node “node” contained in a “network:”
    - Network node
- Two nodes linked together contained in a network
    - Network node1 interaction with node2

2.4.1.3.4. Graph Markup Language (GML)

GML unlike the SIF format comes with a language that supports rich graph formatting and is widely supported by most visualization software tools. A GML formatted file can contain information pertaining to graphs, nodes, and edges, and hence capable of emulating almost every other format. A network can be built using the SIF format and by applying network layouts can then be stored as a GML file as this preserves the layout of a network. Further details on the GML specification can be found on the GML documentation website: http://www.fim.uni-passau.de/index.php?id=17297&L=1.

Other formats such as XGMML is the XML extension of the GML format and is the preferred format to GML, Systems Biology Markup Language (SMBL) format is an XML format used to describe biochemical networks, the specification for SMBL can be found on the website: http://sbml.org/Documents/Specifications, PSI-ML format specification is an XML-based format that is used for data exchange of protein-protein interactions. GraphML is another XML-based format for generating graphs. Apart from the XML-based formats, JSON-based file formats are increasingly being used for data exchange of protein-protein interactions.2.3.

2.4.1.4. Network analysis and visualisation tools

This section discusses some of the commonly used tools in the proteomics network analysis, but before delving into what tools to use, we begin this discussion by looking at the ways by which networks can be quantified in order to provide more informative results.

2.4.1.4.1. Quantifying networks

The most commonly applied metric are; degree, degree distribution, scale-free networks, the degree exponent, shortest path, mean path length, and clustering coefficient Barabasi and Oltvai [107]. By using these network metrics, we can quantify and characterise important network features which are not commonly visible.

Protein-protein interactions are the most commonly used form of networks in proteomic data analysis. In these networks, proteins are represented as nodes while interactions between the nodes are depicted by edges or links. This mapping of proteins is based on experimental information which has been obtained from methods such as mass spectrometer [108], protein chip technologies [109, 110], yeast two-hybrid screens [111], and predictions from computational methods [112]. These mappings have been collected and deposited into online databases as discussed below.

Network tools are mainly used to analyse proteomic data through functional annotation, knowledge integration, modularity analysis, topological analysis and basic network property analysis [113].

The basic properties of a network such as; node degree, degree distribution, betweenness centrality and eigenvector centrality can be used to deduce the significance of a protein [114]. Another important metric is the identification of modules which represent a vital level of organisation in biology [115]. A module in proteomics can be defined as a set of interacting proteins that can be associated with a common biological process. By using networks, clusters of interacting proteins can be identified as modules and associated with a functionality. Modules provide a comprehensive and global description of interaction patterns to comprehend the complexity of biological systems [116]. Module detection enables functional annotation of constituent proteins and the discovery of targets for therapy in diseases such as cancer. In addition to detection of modules, the integration of existing knowledge into networks plays a vital role in the analysis of proteomic data. Such knowledge may include integrating Gene Ontology (GO) annotations, differential gene expression, and pathway details. By highlighting such information, candidate disease proteins may be identified and module functions can be annotated.

2.4.1.4.2. Steps to performing network analysis

To perform network analysis on proteomic data, there are a number of steps that are involved, these steps are summarized in Figure 2-7**.** The steps involved include but are not limited to;

1. The first step involves identifying a list of proteins or genes that need to be analysed using a network tool. The researcher can select which protein or gene appears on the lists, as per individual needs.
2. Interacting partners of these proteins are then obtained from any of the databases discussed above.
3. A protein-protein interaction network is then built by using a visualizing tool from the tools listed in Table 2-2.
4. To get more meaningful information from the network, the protein-protein interaction network is then integrated with already existing knowledge such as pathways, differential expressions for genes or proteins obtained from either high throughput custom data or online databases such as The Cancer Genome Atlas (TCGA). Other existing knowledge that can be integrated includes Gene Ontology enrichment, which can help to identify the functional annotations of the modules or individual proteins in the network.
5. During topological analysis, network theory concepts such as degree, degree centrality distribution, Eigenvectors, degree distribution etc. are applied to identify proteins or nodes playing significant roles in the network, variations between a normal and an altered network and modules that can be mapped to a functionality.
6. Topology analysis is further followed by downstream analysis whose objective is mostly dependent on the researcher.
7. Some of the results that may be obtained from a network analysis of proteomic data include a visual representation of the network, module identification, network variations as well as functional enrichment of proteins and modules.

Table 2-2: Summary of Network tools for analysing proteomic data

Tool	Reference	URL link	Features
Cytoscape	[106]	http://cytoscape.org/	- Open source - Data integration - Network visualisation - Network Analysis - Functional enrichment - Extensible by plugins - Standalone - Platform independent
FunRich (Functional Enrichment Analysis)	[93]	http://funrich.org/	- Open source - Functional enrichment - Dataset comparison - Network visualisation and analysis - Standalone - Runs only on Windows - Results can be exported in various formats
MetaCore	By Thomson Reuters	https://portal.genego.com/	- Proprietary

Tool	Reference	URL link	Features
			- Network visualisation - Network analysis - Function enrichment analysis - Data mining toolkit - Network alignment
Ingenuity Pathways Analysis	IPA®, QIAGEN Redwood City	www.qiagen.com/ingenuity	- Proprietary - Network visualisation and Modelling - Causal network analysis - Network analysis - Functional enrichment analysis - Pathway enrichment analysis - Literature mining - Allows for collaboration
Gephi	Gephi	https://gephi.org	- Network visualisation - Network analysis - Network clustering - Module identification - Dynamic network analysis

Tool	Reference	URL link	Features
			- Real-time visualisation
PINA: Protein Interaction Analysis	[117]	http://cbg.garvan.unsw.edu.au/pina/	- Network construction - Module detection - Functional enrichment - Network metric analysis - Network visualisation - Community drove annotation
Osprey	[118]	http://biodata.mshri.on.ca/osprey/servlet/Index	- Network visualisation - Integrates BioGRID - Ability to compare functions between datasets, - Build interaction network from custom datasets, - Search for specific genes within a network - Filtering feature

2.4.1.4.3. Cytoscape

Cytoscape developed by Trey Ideker (a leading pioneer of systems biology) is a platform independent and open source software tool for the integration, visualisation and statistical modelling of molecular networks together with other systems-level data [105, 119]. The core of Cytoscape provides users with the fundamental features to perform functions such as data integration, analysis, and network visualization. The core also has limited information stored but interconnects with other databases to obtain relevant information. Cytoscape functionality is extensible through the integration of plugins (http://apps.cytoscape.org/) which are now called apps from version 3.0 of Cytoscape.

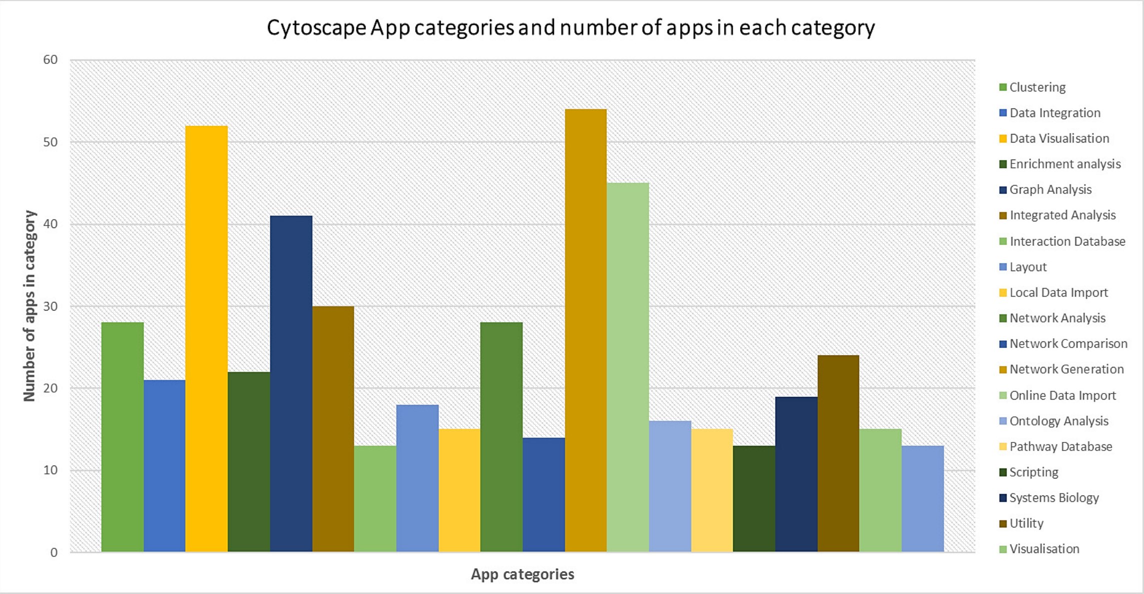


Figure 2-8: The distribution of apps or plugins across a number of categories in Cytoscape.

The apps can be categorised into one or more of the following functional categories such as clustering, data integration, data visualization, enrichment analysis, graph analysis, and integrated analysis. Other functional categories include; interaction database, layout, local data import, network analysis, network comparison, network generation, online data import, ontology analysis, pathway database, scripting, systems biology, utility, and visualization. Figure 2-8 shows the distribution of these apps across the different functional categories.

The first step to a typical Cytoscape workflow is the importation of interactions. These interactions are imported from either a user's own experiment data or from public databases. Data from experiments is loaded directly into Cytoscape using a standard file format such as generic tabular formats including CSV, Excel, and TSV or network-specific formats such as SIF, XGMML, GML, PSI-MI, BioPAX (Biological Pathway Exchange), OpenBEL (Open Biological Expression Language) and SBML.

Importation of data from databases, on the other hand, requires the installation of plugins (apps). A list of genes of interest is passed as a query for interactions from the database. Examples of apps for importing data from databases is the BioGRID database plugin that can be used to import an entire interactome from the BioGRID database. Other ways in which networks can be imported into a network by mining interactions directly from literature or using computational inference from non-interaction data such as expression profiles. This is also achieved through the use of third-party apps. An example of such apps that is Agilent Literature Search software which is a meta-search tool that can automatically search through multiple texts based search engines to extract associations among a set of genes or proteins of interest.

Once the networks are imported into Cytoscape and network visualisation is done, network analysis is achieved using the huge collection of apps. For example, using network topology apps like Knowledge-fused Differential Dependency Network (KDDN), users are able to calculate such statistics as network distribution of node degrees. Network clustering apps such as MCODE enable users to extract network regions which are densely connected thereby forming modules which can then be related to complexes or pathways. Network enrichment apps are used to infer the functions of the identified modules by detecting functional terms that are statistically overrepresented among the set of genes making up the module. Examples of apps that can perform functional enrichment include; BiNGO which is a tool that can determine which Gene Ontology categories are statistically overrepresented in a set of genes or a module, the ReactomeFIPlugin is another app that can be used to associate a set of genes in a module to pathways that are related to diseases such as cancer. Furthermore, functional modules can also be identified by integrating networks with expression data to infer network regions that are consistently up- or downregulated. Another example of network analysis that can be done using apps in Cytoscape is network comparison, this involves comparing networks across species or in different conditions to identify regions of the network with conserved interactions. GASOLINE (Greedy and Stochastic algorithm for Optimal Local Alignment of Interaction NEtworks) is an example of an app that can be used to compare multiple networks.

Cytoscape also supports the use of scripting languages such as Python and R. It enables users to develop their own scripts and integrate or call Cytoscape functionality in the order they want it to be done.

2.4.1.4.4. FunRich

Functional Enrichment Analysis (FunRich) tool [93] is an open source standalone desktop software tool for functional enrichment and protein-protein interaction network analysis of biological molecules. Features of FunRich include functional enrichment and network analysis of genes and proteins. In addition, FunRich allows the representation of results in editable graphical form as Venn, Bar, Column, Pie and Doughnut charts. FunRich users can perform a biological process, cellular component, molecular function, protein domain, site of expression, biological pathway, transcription and clinical synopsis phenotypic term enrichment. Users can analyse their datasets against two built-in background databases; FunRich and UniProt or against a customized background database. FunRich does not require users to install any additional applications or plugins to conduct any of the above analysis. FunRich is currently only available for Microsoft's Windows Operating system with plans underway to support other major operating system platforms.

The first step to performing an enrichment analysis in FunRich is the specification of an annotation database. By default, FunRich comes with a human annotation database. Each database consists of biological function annotations and an interaction database. FunRich also comes with the latest UniProt annotation database, otherwise, users can also include a custom database. Once an annotation database has been specified, a list of genes or proteins is then imported. The user can perform a range of analyses on the datasets including comparison across the datasets using a Venn diagram that shows which proteins or genes are common across the datasets. Users can also perform gene set enrichment analysis to determine what biological functions are statically enriched in the gene or protein lists. In addition to these, FunRich also allows users to generate and build an interaction network from where users can then manipulate the network through enriched pathways and functions.

2.4.1.4.5. MetaCore

MetaCore from Thomson Reuters is an integrated proprietary software suite capable of analysing multiple types of biological data, for example, Next Generation Sequencing [120], variant, Copy Number Variation (CNV), microarray, metabolic, proteomics, microRNA etc. Functional analysis in MetaCore is performed against a high quality, a manually-curated database containing molecular interactions vis-à-vis protein-protein interactions, protein-DNA interactions, and protein-RNA interactions. The database is also made up of molecular classes such as transcription factors, signalling and metabolic pathways and disease ontologies. MetaCore was developed for the purpose of representing biological functionality along with the integration of functional, molecular, or clinical information. Using the data mining toolkit available in MetaCore, users can perform functions like data visualization, analysis and exchange of data, network alignment using multiple network alignment algorithms, and enrichment analysis. While MetaCore provides a set of rich features, it is a paid for a suite of software for integrated analysis.

2.4.1.4.6. Ingenuity Pathways Analysis

IPA (IPA®, QIAGEN Redwood City, www.qiagen.com/ingenuity) is a proprietary software application with features that allow scientists to model, analyse and understand the complexity of biological and chemical systems [121]. IPA offers a host of network analysis functions some of these include; causal network analysis allows researchers to identify upstream molecules that control the expression of genes in their datasets and network analysis which allows the building and exploration of transcription of molecular networks such as microRNA, transcriptional networks, and Protein-Protein interaction networks. Network analysis in IPA can identify regulatory events that lead from signalling events to transcriptional effects, help in understanding toxicity responses by exploring connections between drugs or targets and related genes or chemicals. Users can also edit and expand networks based on the molecular relationships most relevant to the project.

IPA is capable of identifying pathways, molecular mechanisms and biological processes that are relevant to a given dataset. It is also capable of finding biological and chemical knowledge from the scientific literature. Other features allow for collaboration, sharing of results and insights with project teams.

IPA is a subscription-based software application. It is made available as a web-based, hosted or deployed solution.

2.4.1.4.7. Gephi

Gephi is an open-source data exploratory, network visualization and analysis software tool for large network graphs. Gephi allows users to explore, analyse, spatialize, filter, cluster, manipulate and export all types of network graphs. With Gephi, users can derive hypotheses and identify patterns by analysing data using networks.

Gephi can be used to analyse a variety of networks ranging from biological networks to social networks. It supports the majority of the network file formats discussed in section 2.2 above. The core of Gephi can perform basic network metric analysis such as calculating betweenness centrality, closeness, clustering, community detection or module identification. Gephi further includes a feature that allows for the analysis of dynamic networks were a set of networks representing or derived from different conditions or events are compared to infer differences. In addition, Gephi is also extensible by a range of plugins which users can install to perform functionality that is not included in the core of Gephi. While Gephi provides a range of network analysis features, other biological specific network analysis features such as functional enrichment cannot be easily done due to the unavailability of such functionality within Gephi or its associated plugins.

2.4.1.4.8. NDEx-The Network Data Exchange

NDEx-The Network Data Exchange is not so much a network analysis tool, but rather an open source framework for sharing of networks of many types and formats, publication of networks as data, and the use of networks in modular software [122]. Unlike other similar tools such as KEGG and IntAct, NDEx is a data commons framework that allows users to manage the sharing and the publication of networks. Users can upload any type of networks such as pathway models, interaction maps, and novel data-driven knowledge networks. NDEx supports networks of varying formats including simple interaction format (SIF), extensible graph markup and modelling language (XGMML), BioPAX3 and OpenBEL. Each network uploaded to NDEx is given an accession number which acts as a universally unique identifier allowing users to share or include such networks in publications. NDEx also promotes the development of network analysis algorithms and applications by providing access to networks which can be used as inputs through a web-based relational state transfer application programming interface (REST API). In addition, users can anonymously access networks by searching through the web interface (www.ndexbio.org). The framework can also be downloaded and run on a local server or personal computer, depending on the needs of a user.

2.4.1.4.9. PINA: Protein Interaction Analysis

Protein Interaction Analysis is a web-based integrated network analysis platform for protein interaction network construction, filtering, analysis, visualization and management [117]. PINA has a quarterly updated backend database consisting of an integration of data from six other publicly available databases; IntAct, MINT, BioGRID, DIP, HPRD and MIPS MPact. To construct a network, PINA provides a query feature where users can either query a single protein, a list of proteins, a list of protein pairs or two lists of proteins.

The constructed PPI networks can be further analysed by PINA's inbuilt GO term and protein domain annotation tools. Other analyses that can be performed include the use of graph theoretical tools to either discover basic topology properties of a PPI network or identify topologically important proteins, such as hubs or bottlenecks, based on several centrality measures from protein domains and GO terms. In addition, the constructed networks can be downloaded in customized tab delimited, GraphML or MITAB formats for further analysis using tools such as Cytoscape where they can be integrated with gene expression profiles.

2.4.1.4.10. Colorectal Cancer Atlas

Colorectal Cancer Atlas [11] is an integrated web-based resource mainly meant for those involved in colorectal cancer research. The tool provides a platform that catalogues both non-quantitative and quantitative proteomic and genomic sequence variation data in both colorectal cancer cell lines and tissues. This information has been curated from existing literature.

Colorectal Cancer Atlas features an easy to use search functionality that also offers auto-complete. Users can search for a given protein, gene, pathway or cell line that may be of interest to them. Depending the type of search term, the tool then performs functional, pathway and GO enrichment, maps sequence variances known in colorectal cancer and associated with the searched term, and generates a protein-protein interaction network.

The network integrates proteomic data with genomic sequence variations. Users can use this network analysis module to quickly get an overall picture of the interacting partners of a given gene in colorectal cancer. It uses colour intensities to indicate the number of sequence variances for a given gene in the database. Users can also filter through the network by either a gene symbol or by cell lines.

While this tool is specific to colorectal cancer, it provides features that users can quickly use to get functional enrichment information for a given protein or gene as well perform a gene or protein centred network analysis. Overall, researchers can quickly look up a list of genes or proteins and get an overview of a given gene in colorectal cancer.

2.4.1.4.11. Osprey

Osprey [118] is a software tool that allows for the visualization and analysis of complex interaction networks. Just like most visualization tools, in osprey genes are represented as nodes and interactions as edges. Developed using Java, Osprey is platform independent running on both Linux and Windows based systems.

Osprey provides a range of features that allows users to easily build data-rich graphical representations of their datasets. In addition, users can use the default BioGRID's Gene Ontology interaction datasets to quickly build an interaction network. Some of the features in Osprey include; ability to compare functions between datasets, use of custom datasets to build interaction networks, ability to search for specific genes within a network as well filter functions to filter for specific nodes within a large a network. Osprey also has a number of network layouts including concentric circles, spoke, circular and dual ring, these layouts allow for the comparison of large-scale datasets in an additive manner.

2.4.1.5. Conclusions

In order to study and understand complex systems such as cellular systems, we have shown that network theory provides metrics that can be used to study such systems using a bottom up approach. In this chapter, we have given an overview of how network theory can be applied to the analysis and study of proteomics data based on a number of network theory metrics. Such metrics include; node degree, node centrality, Eigen vector values and modularity.

We have also discussed the most frequently used network analysis tools in analysing proteomic data. In doing so, a generic workflow that one can use during the analysis has been described. Tools discussed included databases which are used to house protein-protein interaction network annotations and the analytical tools that can be applied in analysing proteomic data.

2.4.2. Computational analysis of PPIs in disease

Section 2.3.1 discussed the pivotal roles played by PPIs in pathological states. Therefore, understanding the interactions among proteins is vital in the inference of proteins and modules responsible for tumour progression and metastasis in cancer. Against this background and given the cost of conducting wet laboratory-based experiments, several computational tools and methods have been developed and widely applied in analysing PPI networks to identify genes, together with their corresponding proteins and protein modules involved in cancer progression. For instance, tumour related genes and protein networks are inferred by computationally integrating PPI networks with gene expression data from tumours [44, 123, 124]. Sun and Zhao [125], on the other hand, found network topological differences for proteins encoded by known cancer genes upon analysing their global and local network characteristics. Interestingly, Yang, et al. [126] using gene co-expression networks showed that proteins that are encoded by cancer prognostic genes do not generally form hubs (proteins that are highly connected) within PPI networks, but are often found enriched in modules (groups of highly interconnected proteins) that are highly conserved. Similarly, Brown, et al. [127] analysed gene co-expression networks in glioblastoma and identified CD133 and CD44 genes as indicators of the different glioblastoma subtypes based on their modules of enrichment.

One of the principal areas of active research in biomarker discovery is the prioritisation of genes from among thousands of other candidate genes. High-throughput (HTP) techniques such as linkage analysis [128] and GWAS [129] are typically used in associating genetic variations to diseases [130]. While the cost of conducting such types of studies has been decreasing over the years with the advent of new technologies, doing so is, however, time consuming due to the increased data amounts and is also prone to false positives [131]. As such, the development of computational methods and algorithms to comprehensively prioritise such candidates before wet laboratory experiments are conducted can help reduce such costs [132]. Gene prioritisation involves the assignment of confidence scores to genes based on the probability of being associated with a disease [133]. Several bioinformatics tools have thus been developed to identify genes associated with a disease by combining various data sources such as PPIs, gene expression, functional similarities, and pathway annotations. Perez-Iratxeta, et al. [133] developed the first major type of such tools, and since then, other numerous tools and algorithms have been implemented [134-141].

Interestingly, the common underlying theme among these methods is that they are often based on the principle of "guilty by association" or GBA. That is, genes or proteins that are similar to or interact with genes/proteins that are already known to be associated with a disease are then more likely to be functionally related or associated with the same disease [40, 142-145]. That is, two proteins that are closer to each other in a network are bound to be functionally similar. Using this principle, basic network metrics are used to determine the distance between two proteins in a network. Examples of such metrics include shortest-path where the shortest distance between two proteins is the lowest number of edges connecting the two proteins. Others include the diameter, neighbourhood, clique, cut, node degree, and density [90, 146, 147]. Computational GBA methods have been applied widely to infer novel protein functionality as well as associating genes to diseases, for instance, Wolfe, et al. [142], by analysing co-expression networks, found genes with similar functionality as other already known genes. Similarly, Wu, et al. [148] developed a tool called CIPHER that predicts and prioritises genes associated with disease, and Zhou, et al. [149], on the other hand, incorporated biomedical literature to the development of a symptom-based human disease network and found that the similarity of symptoms between two diseases correlated with the number of genes associated with both diseases as well as the extent to which their related proteins interact, thereby showing that proteins that are related participate in similar phenotypes.

To generalise beyond the direct interacting neighbours, other methods have also been developed. Such methods include module-based methods which first group or cluster together a group of related proteins and infer the function of the module, based on the function of the members [150, 151]. In such methods, statistical and machine learning techniques are used to group similar proteins based on a wide range of features [147, 152-161]. For example, Menche, et al. [162] established a set of mathematical conditions that showed that diseases with overlapping modules had statistically significant molecular similarity. Nonetheless, using such methods has proven to be ineffective and inconsistent in linking the functional roles of proteins against several phenotypes compared to GBA based methods[138, 156, 163, 164].

Furthermore, to address some of the shortcomings of modular-based methods, recent methods have been proposed that take into consideration the global topology of a PPI network when inferring disease-associated genes [138, 164]. At the core of these methods is the concept of network propagation, a network analysis technique whereby a biological signal is broadcast through the entire network [165, 166]. The biological signal is amplified accordingly by those proteins that are considered to be functionally related to the protein that generated the signal [164]. Biological signals, in this case, can be prior information that associates genes with a given phenotype, such as a disease like CRC. Network propagation is performed by first overlaying the prior information on network nodes (proteins). This information is then propagated from each node across the network via the edges to neighbouring nodes repeatedly until the number of steps specified is reached or upon convergence [166]. The final scores of each node are therefore dependent on the scores of its interacting partners whose scores are also dependent on their neighbours, and so forth.

Network propagation, therefore, provides research scientists with the opportunity to integrate networks with various types of heterogeneous data. Network propagation has long been applied in several areas of scientific research while taking on different forms and names [164, 167-171]. In systems biology, network propagation has also been used in such areas as gene-disease association studies [76, 172], module detection [156, 173], gene function characterisation [151, 174] as well as in the discovery of drug targets [175].

2.5. Exosomes and exosome biogenesis

Cancer is a highly complex and heterogeneous disease which is sustained by a robust biological system of networks such that they gain the ability to survive, adapt and proliferate even in the presence of anticancer drugs [176-178]. The biological networks are derived from interactions between proteins, genes, DNA, RNA, and other small molecules within cells, as well as intercellular and distant cell interactions [177]. For the biological system of networks to be sustained, there is a need for the influx and efflux of biological materials across the nuclear and plasma membranes [179]. It is well-established that cellular systems consist of active and passive transport machinery which handle the movement of biological materials in and out of cells via the membranes [180, 181]. However, recently, research has shown that there are other transport mechanisms such as extracellular vesicles (EVs) which are involved in both short and distant intercellular communication [182, 183].

Over the last decade, research has further shown that EVs transport several active biological cargoes such as DNA [184], proteins [185], RNA [186], lipids [187], viruses [188], and metabolites [189]. Also, the transported cargo is reflective of their cellular origin and is capable of affecting the recipient cell's phenotype [183, 190]. EVs, through the transfer of their cargo, can regulate several biological functions ranging from normal physiological processes such as cell maintenance and tissue repair [191] to pathological processes that underlie diseases such as cancer [182]. There are three main categories of EVs based on their biogenesis [183]: exosomes, ectosomes or shedding microvesicles (SMVs) [183, 192] and apoptotic bodies (ABs) [193]. In this thesis, we focus on exosomes.

Exosomes are 30-150 nm in diameter membranous vesicles of endocytic origin that are secreted by a variety of cells under normal and pathological conditions [183, 194]. First reported by Pan and Johnstone [185], exosomes are bound by a lipid bilayer membrane that encloses a small cytosol but lacks organelles. Like other EVs, exosomes contain various biological materials such as proteins and nucleic acid materials that are reflective of their cell of origin [183, 195]. While the content of exosomes varies according to their cell of origin, research has shown that they contain a set of protein molecules which are evolutionary-conserved [196]. Over the years, there has been increased interest in the role of exosomes in both physiological and pathological conditions due in part to their ability to carry biological content between cells. The role of exosomes in physiological conditions is poorly understood while ongoing research has implicated exosomes in several pathological conditions such as cancer where it is shown that exosomes are involved in the metastasis of cancer [197-199], drug resistance [177], and epithelial-to-mesenchymal transition (EMT) [200]. On the other hand, exosomes are proposed as potent vehicles for the delivery of therapeutic drugs [201, 202]. In addition, because exosomes are secreted into readily available body fluids such as blood and urine, they can be used as biomarkers for the diagnosis and prognosis of cancerous tumours [203]. Because of this enormous potential that can be harnessed by understanding the role of exosomes in both physiological and pathological states, there has been a growing interest in the study of the biogenesis, functions, and applications of exosomes.

The biogenesis of exosomes starts with the inward budding of endosomal membranes which results in the formation of intraluminal vesicles (ILVs) within the multivesicular bodies (MVBs) [204]. Upon maturation, the MVBs fuse with the plasma membrane and their contents are then secreted into the extracellular space as exosomes, as summarised by the flowchart in Figure 2-9. However, the mechanism behind exosome biogenesis is still poorly understood. Nonetheless, the endosomal sorting complex required for transport machinery together with other accessory proteins is thought to be one of the mechanisms by which exosome biogenesis is regulated [191, 205].

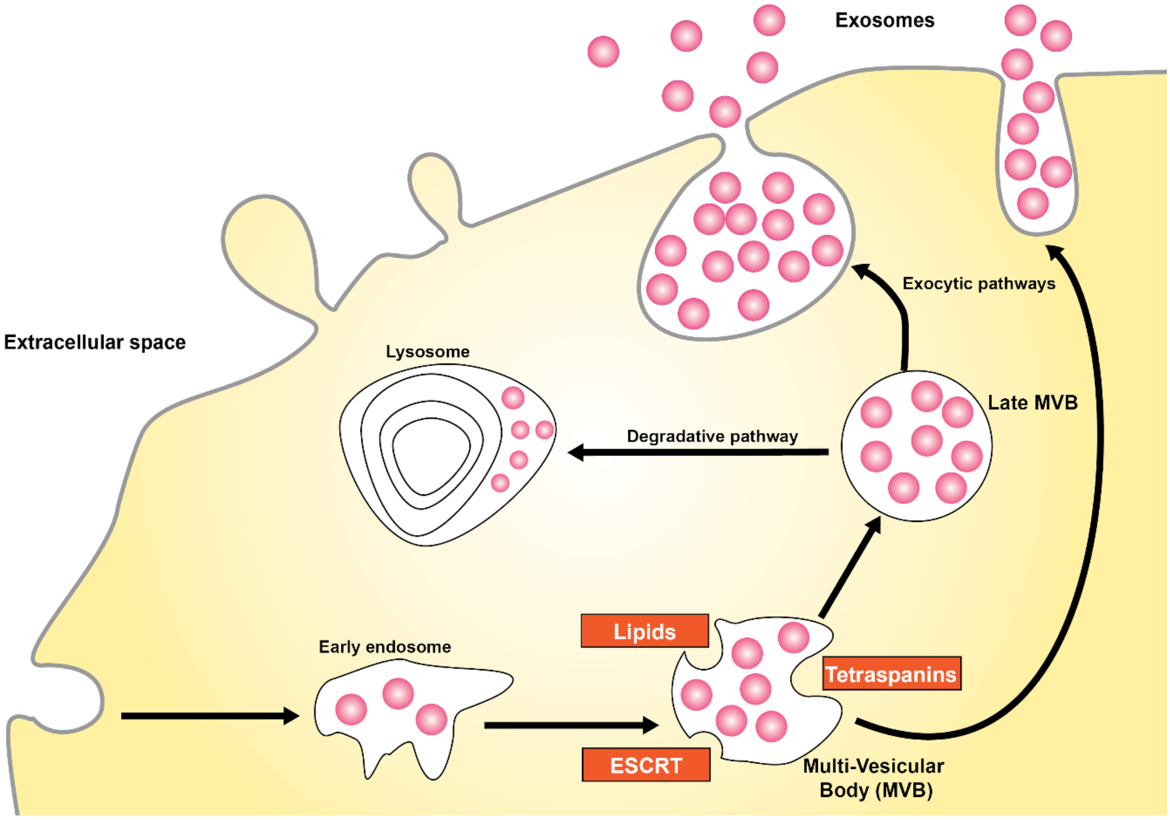


Figure 2-9: Exosome biogenesis and release. Biogenesis of exosomes starts with the inward budding of endosomal membranes which results in the formation of intraluminal vesicles (ILVs) within the multivesicular bodies (MVBs). Once the MVBs mature, they fuse with the plasma membrane and their contents are then secreted into the extracellular space as exosomes. Several mechanisms are implicated as being involved in this process: the ESCRT machinery, tetraspanins, lipids and Rab GTPases. The mechanisms by which they achieve biogenesis are, however, still poorly understood.

The ESCRT machinery consists of about 20 proteins divided among four complexes (ESCRT-0, -I, -II, and -III) as shown in Figure 2-10. The four ESCRT components are linked together and work in a sequential order in partnership with other accessory proteins (such as ALIX, VPS4 and VTA1) to regulate exosome biogenesis [204]. According to Schmidt and Teis [206], the ESCRT-0 complex initiates the process of MVB formation by first localising to endosomes where it binds to degradation-bound ubiquitylated proteins, hence making it the first step i sorting in the MVB pathway. It is made up of the proteins

STAM, STAM2 and HGS. The ESCRT-0 complex also recruits the ESCRT-I complex when the HGS subunit binds to ESCRT-I's TSG101 subunit. The ESCRT-I complex forms a rod-like shaped complex of proteins consisting of UBAP1, VPS37(A-D), TSG101, VPS28, MVB12 (A and B) where one end of the complex with TSG101 binds to the ESCRT-0 complex and other ubiquitylated proteins, thereby forming the second phase of sorting. The ESCRT-I complex also works together with the ESCRT-II complex to initiate the inward budding process of the endosomal membrane. The ESCRT-II complex is a Y-shaped complex made up of the proteins SNF8, VPS25, and VPS36 with VPS36, forming the hub-base of the complex that binds to ESCRT-I as well other proteins such as PI3P and ubiquitin [191, 206].

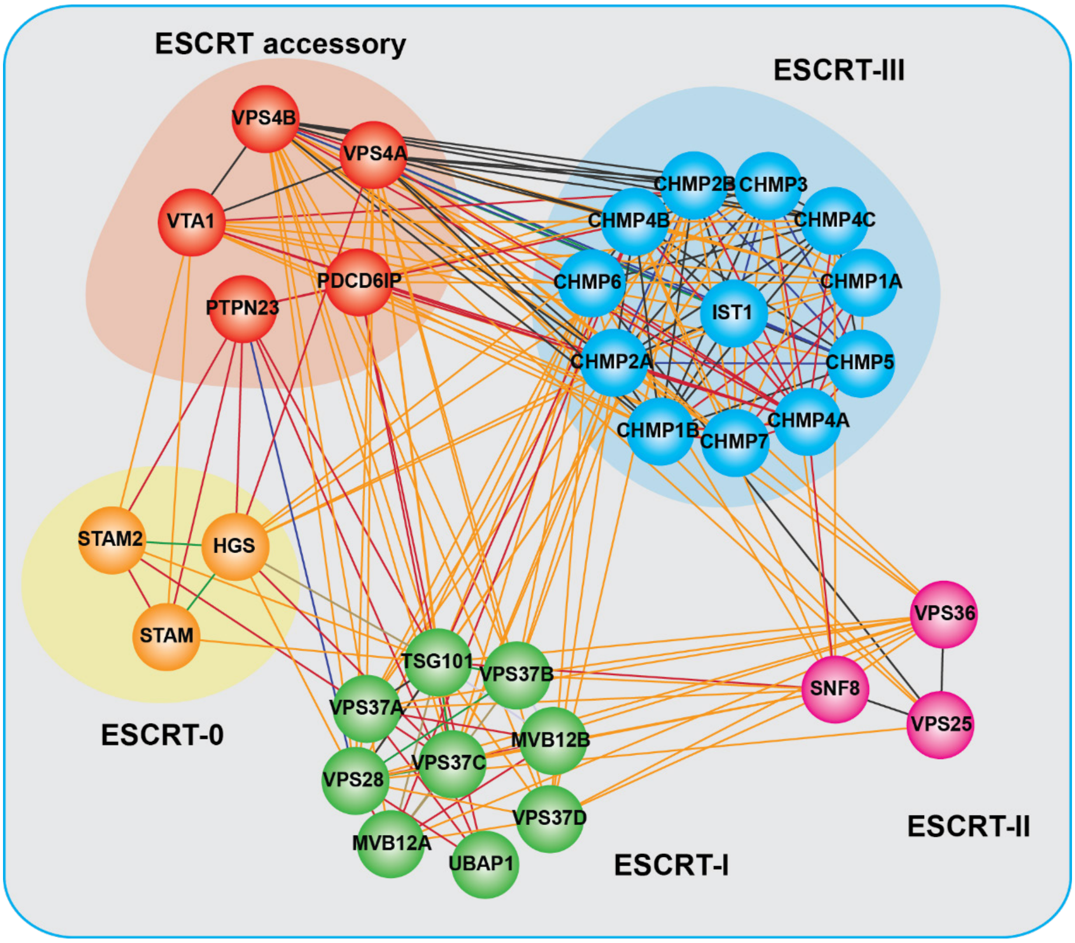


Figure 2-10: ESCRT machinery complexes. The four complexes (ESCRT-0, I, II, III) that make an ESCRT machinery together with their associated accessory proteins.

Therefore, the first three ESCRT complexes (ESCRT-0, I and II) are said to work together by binding to ubiquitinated cargo and sorting ubiquitinated membrane proteins into MVBs during exosome biogenesis. The ESCRT-III complex on the other hand is made up of the proteins CHMP1(A, B), CHMP2(A, B), CHMP4 (A-C), IST1, CHMP (5-7). The ESCRT-III complex is responsible for the sequestration of cargo within vesicles and membrane budding [206].

Other than the ESCRT machinery, recent studies have shown that other ESCRT-independent pathways are likely to regulate exosome biogenesis. For instance, a study by Stuffers, et al. [207] showed that in the absence of essential ESCRT proteins in mammalian cells, the formation of MVBs and the secretion of exosomes were not wholly impaired which, therefore, implies that other pathways exist that regulate exosome biogenesis. Examples of such ESCRT-independent pathways include tetraspanins-enriched domains [208], Rab GTPases [209] and lipids [210]. Interestingly, the regulation of exosome biogenesis by the ESCRT-dependent pathway, unlike their counterparts, has been shown to be conserved in other species [210, 211]. Despite the mounting evidence of the role of ESCRT machinery as well as other pathways in exosome biogenesis, the biology and the mechanisms behind exosome biogenesis are yet to be fully understood.

2.6. Conclusion

The computational analysis of PPIs has contributed immensely to the advancement of our knowledge of the changes in the biological systems of diseased states. However, to further build on what is already established, there is need to develop novel methods that can integrate both clinical and molecular data to infer new clinical phenotypes that are relevant in areas such as personalised medicine [145]. Over the years, there has been a transition from traditional bioinformatics to translational bioinformatics.

With the continued advancements in high-throughput data collection techniques, the challenge therefore for bioinformaticians and computational biologists is developing computational tools and methods that are scalable and capable of integrating various heterogeneous data. Hence, today's computational methods for the analysis of PPIs should be capable of integrating environmental factors as well as analysing interactions between different organisms such as host-pathogen interactions. They should also be capable of discovering disease biomarkers through the analysis of PPIs which can ultimately lead to drug discovery. Nevertheless, traditional methods for the analysis of biological interactomes use static networks which do not address several factors which need to be taken into account when analysing biological interactomes. Such factors include biological functions, being time-sensitive, proteins and the fact that the networks they form do not always exist at the same time. Also, biological networks are dynamic in nature; a single protein can serve multiple functions and at the same time can interact with proteins that function completely different from its own. A static network will therefore not account for the spatial and temporal aspects of biological interactomes and may lead to the inaccurate representation of the dynamism that is characteristic of biological networks. To therefore correctly analyse proteomics data in dynamic diseases such as cancer, there is the need to develop computational methods and tools that can encompass the temporal aspects underlying such diseases.

Chapter 3

Colorectal cancer atlas: An integrative resource for genomic and proteomic annotations from colorectal cancer cell lines and tissues

This chapter has been peer-reviewed and published in the Journal of *Nucleic Acids Research* (Chisanga, et al. [11]) and is presented here as a manuscript.

The candidate designed and developed the resource. The candidate was also involved in the collation and annotation of the data as well as the bioinformatics analysis of mass spectrometry data.

David Chisanga[1], Shivakumar Keerthikumar[2], Mohashin Pathan[2], Dinuka Ariyaratne[2], Hina Kalra[2], Stephanie Boukouris[2], Nidhi Mathew Abraham[2], Haidar Al Saffar[2], Lahiru Gangoda[2], Ching-Seng Ang[3], Oliver M. Sieber[4,5], John M. Mariadason[6], Ramanuj Dasgupta[7], Naveen Chilamkurti[1] and Suresh Mathivanan[2,*]

[1]Department of Computer Science and Information Technology, La Trobe University, Bundoora, Victoria, 3086, Australia

[2]Department of Biochemistry and Genetics, La Trobe Institute for Molecular Science, La Trobe University, Melbourne, Victoria, 3086, Australia

[3]The Bio21 Molecular Science and Biotechnology Institute, University of Melbourne, Parkville, Victoria 3010, Australia

[4]Walter and Eliza Hall Institute of Medical Research, Parkville, Victoria 3052, Australia

[5]Faculty of Medicine, Dentistry and Health Sciences, Department of Medical Biology, University of Melbourne, Parkville, Victoria 3052, Australia

[6]Olivia Newton John Cancer Research Institute, Melbourne, Australia, Ludwig Institute for Cancer Research, Melbourne-Austin Branch, Australia, School of Cancer Medicine, La Trobe University, Melbourne, Australia

[7]Genome Institute of Singapore, A*STAR, 60 Biopolis Street, Singapore 138672

*To whom correspondence should be addressed

Dr. Suresh Mathivanan

Department of Biochemistry,

La Trobe Institute for Molecular Science, La Trobe University,

Bundoora, Victoria 3086, Australia

Tel: +61 03 9479 2565

Email: S.Mathivanan@latrobe.edu.au

3.1. Abstract

In order to advance our understanding of colorectal cancer (CRC) development and progression, biomedical researchers have generated large amounts of omics data from CRC patient samples and representative cell lines. However, these data are deposited in various repositories or in supplementary tables. A database which integrates data from heterogeneous resources and enables analysis of the multidimensional datasets, specifically pertaining to CRC is currently lacking. Here, we have developed Colorectal Cancer Atlas (http://www.colonatlas.org), an integrated web-based resource that catalogues the genomic and proteomic annotations identified in CRC tissues and cell lines. The data catalogued to-date include sequence variations as well as quantitative and non-quantitative protein expression data. The database enables the analysis of these data in the context of signaling pathways, protein-protein interactions, Gene Ontology terms, protein domains and post-translational modifications. Currently, Colorectal Cancer Atlas contains data for >13,711 CRC tissues, >165 CRC cell lines, 62,251 protein identifications, >8.3 million MS/MS spectra, >18,410 genes with sequence variations (404,278 entries) and 351 pathways with sequence variants. Overall, Colorectal Cancer Atlas has been designed to serve as a central resource to facilitate research in CRC.

Keywords: colorectal cancer database, colorectal cancer atlas, proteomics, genomics, bioinformatics, databases

3.2. Introduction

Colorectal cancer (CRC) is the third most common form of cancer and has the fourth highest mortality rate in the world [212]. In order to advance our understanding of the initiation and progression of this disease, biomedical researchers have performed global analyses of the genome, epigenome, transcriptome, proteome and metabolome of CRC patient samples and representative cell lines [213-216]. According to The Cancer Genome Atlas Network [214], APC, TP53, KRAS, PIK3CA, FBXW7, SMAD4, TCF7L2 and NRAS are the most frequently mutated genes in CRC. Identification of these mutations and associated pathways has advanced our understanding of CRC, is enabling the sub-classification of this disease and is unveiling potential new avenues for treatment.

Due to the significant advancements in high-throughput technologies, vast amounts of multidimensional data relevant to the biology of CRC have been generated. To extract meaningful biological insights from these data, researchers previously needed to collate data from a large number of studies. To facilitate this process, a series of databases have been created. For example, cancer gene mutations are currently catalogued in databases including TCGA [214], COSMIC [217], TumorPortal [218], IntOGen [219], Network of Cancer Genes [220] and TSGene [221]. These databases provide valuable information of gene variations for a number of tumour types including CRC, however they are not specifically designed to integrate sequence variations with proteomic data. NetGestal [222] is a web-based framework that allows for integration of OMIC data from multiple species in the context of biological networks [223] and contains data pertaining to human CRC from TCGA. However, there is currently no user-friendly online resource specifically pertaining to CRC which catalogues genomic and proteomic data from literature, databases and TCGA, integrates the sequence variations with protein domain, post-translational modifications and protein-protein interactions.

Here, we describe Colorectal Cancer Atlas (http://www.colonatlas.org), an integrated web-based resource which catalogues genomic and proteomic data from CRC tissues and cell lines. Data catalogued includes; quantitative and non-quantitative protein expression, sequence variations, cellular signaling pathways, protein-protein interactions, Gene Ontology terms, protein domains and post-translational modifications (PTMs). Data pertaining to genomic sequence variations and protein expression have been manually curated from the scientific literature and collated from other publicly available databases.

Colorectal Cancer Atlas is designed to enable a user to search for a specific mutation in any particular cell line, and search for cell lines with and without specific mutations. Currently, Colorectal Cancer Atlas contains data for >13,711 primary CRC tissues, >165 CRC cell lines, 62,251 protein identifications, >8.3 million MS/MS spectra, >18,410 genes with sequence variations, 404,278 sequence variation entries, 351 pathways with sequence variants, 88,819 PTMs and 253,700 protein-protein interactions (Table 3-1).

Table 3-1: Colorectal Cancer Atlas statistics

Protein entries	62,251
MS/MS spectra	8,378,422
Primary tissues	13,711
Cell lines	179
Genes with sequence variants	19,831
Gene sequence variants	404,278
Pathways with genes having sequence variants	351
Pathways with genes having no sequence variants	1,657
Cell lines with drug sensitivity	27
PTMs	88,819
PTMs affected by sequence variants	1,631
Protein-protein interactions	253,700

3.3. Database architecture and web interface

Colorectal Cancer Atlas is a web-based application developed using Zope2 (version 2.8.7-1), a python-based web framework. The back-end database is MySQL (version 5.0.95), a well-established open source database. The web pages were developed using Hyper Text Mark-up Language (HTML) in combination with JavaScript for front end functionality, while Python (version 2.4.3), a scripting language was used for database connectivity and back-end processing. JavaScript modules include DataTables (version 1.10.4) for the development of interactive data tables, Data-Driven Documents (D3JS) for the development of interactive protein-protein interaction networks, and Highcharts (version 4.1.6) for the development of interactive heat maps and column charts.

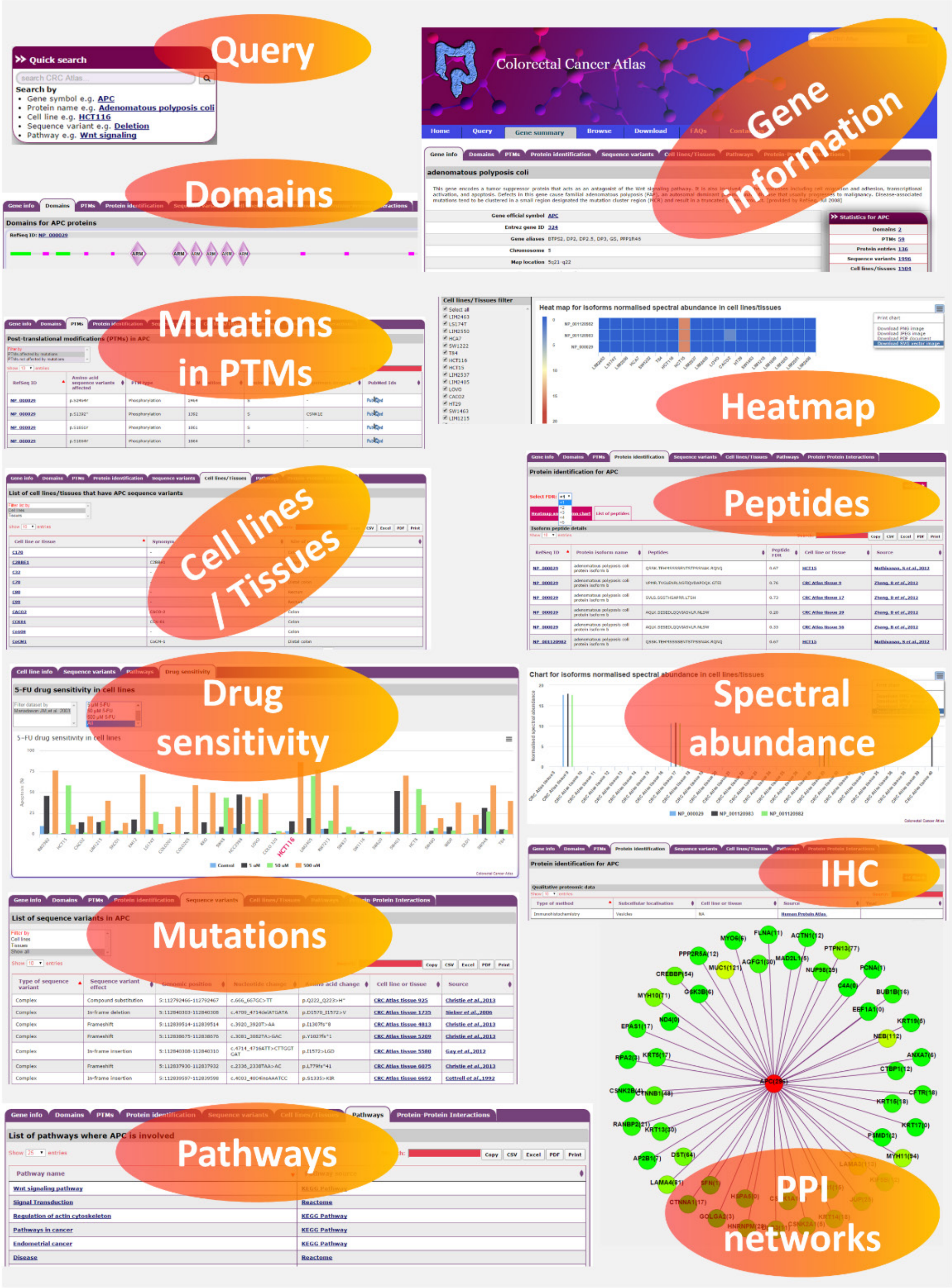
3.4. Genomic datasets

Colorectal Cancer Atlas catalogues gene sequence variations present in primary CRC tissues and cell lines which were collated by manual curation of the scientific literature. In addition, the database contains genomic variations identified in CRC cell lines sequenced in-house. For cell lines, where available, the gender and age of the patient is provided, along with the specific cell type, doubling time, culture properties and stage of cancer. This information was obtained from the Cancer Cell Line Encyclopedia [224], ATCC (http://www.atcc.org), COSMIC database and literature. Sequence variation details including the type of sequence variants, putative mutational effects, nucleotide change and amino acid changes are displayed.

3.5. Proteomic datasets

Colorectal Cancer Atlas also catalogues proteomic data collated from multiple resources including the scientific literature (e.g., Zhang *et al.* [216]), Human Protein Atlas [225], Human Proteinpedia [226] and Human Protein Reference Database [227]. Experimental techniques used in generating these data included mass spectrometry, Western blotting, immunohistochemistry, confocal microscopy, immunoelectron microscopy and fluorescence-activated cell sorting (FACS). In addition, publicly available label–free quantitative mass spectrometry data for CRC cell lines and tissues were re-analysed using an in-house proteomics pipeline in order to provide standardized data. The proteomics pipeline involved conversion of raw mass spectrometry data files into the Mascot Generic File Format (MGF) using MsConvert with peak picking [228]. The MGF files were then searched using X! Tandem (Sledgehammer edition version 2013.09.01.1) [229] against a target and decoy Human RefSeq protein database. Peptides were further filtered using <5% false discovery rate (FDR) as a cut-off, and quantified using the Normalized Spectral Abundance Factor (NSAF) method [230].

3.6. Colorectal Cancer Atlas provides an integrated view of multiple data types



Quick search

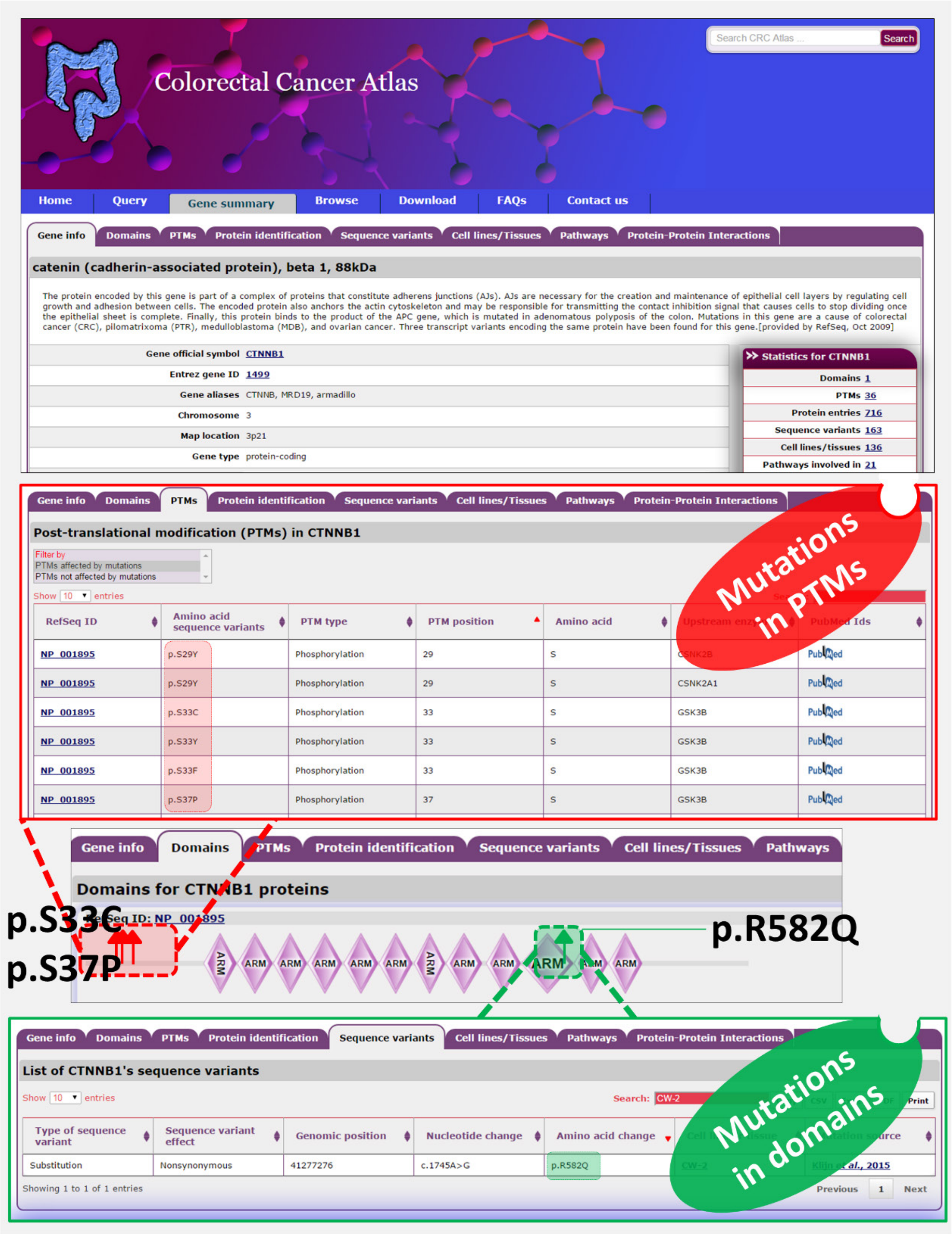
Search by
- Gene symbol e.g. APC
- Protein name e.g. Adenomatous polyposis coli
- Cell line e.g. HCT116
- Sequence variant e.g. Deletion
- Pathway e.g. Wnt signaling

Figure 3-1: Snapshot of Colorectal Cancer Atlas features. An overview of proteomic and genomic data features for the APC gene is displayed. A user can query the database using a gene symbol or a protein name. A gene information page will provide the users with details pertaining to protein domains, post-translational modifications (PTM), reported mutations in cell lines/tissues, quantitative protein expression, pathway, protein-protein interaction (PPI) and cell line drug sensitivity.

Colorectal Cancer Atlas provides an integrated view of the sequence variations and the proteomic data. Mass spectrometry-based quantitative proteomic data are depicted as heat maps and column charts in the respective molecular pages (Figure 3-1), and users are able to filter the datasets based on the FDR. The database also contains protein expression data generated using immunohistochemistry, Western blotting, FACS, confocal and immunoelectron microscopy. The database also includes protein data derived from various cellular fractions including the nucleus, cytoplasm, membrane, the secretome [231] and exosomes [232] (from ExoCarta [233]).

The integration of sequence variants with proteomic data is designed to facilitate the prediction of functional effects of the protein. For each gene, Colorectal Cancer Atlas enables parallel visualization of CRC associated sequence variants with quantitative protein expression across CRC cell lines and tissues. In addition, PTMs, and protein domains affected by the sequence variation can be visualized (Figure 3-1), enabling the potential effect of sequence variants on protein function to be easily ascertained. For example, β-catenin mutations in positions S33, S37, T41 and S45 occur in CRC, all of which are critical for phosphorylation [234]. Mutations in these serine/threonine residues allows for the stabilization of β-catenin and constitutive activation of the Wnt signaling pathway. Similarly, Colorectal Cancer Atlas displays sequence variations in known protein domains which can provide valuable insight into the putative effect on protein function. For example, mutations in the armadillo domain (R582) in β-catenin have been described which have been reported to alter the binding of β-catenin to TCF4 [235] (Figure 3-2).

Colorectal Cancer Atlas also provides a graphical representation of known protein interactions (obtained from BioGrid [236] and Human Protein Resource Database [227]), where each protein is depicted as a node with a specific colour and intensity corresponding to the number of sequence variants in the encoding gene (Figure 3-1). Furthermore, Colorectal Cancer Atlas integrates biological pathways with gene sequence variants. Biological Pathways were obtained from Reactome [237], KEGG [238], Cell map and HumanCyc. For example, as shown in Figure 3-1, sequence variants in APC are implicated in the dysregulation of the Wnt signaling pathway and actin cytoskeletal remodelling. Finally, Colorectal Cancer Atlas contains data on 5-flurouracil (5-FU) drug sensitivity for CRC cell lines curated from the literature (studies using at least 3 CRC cell lines [239]). Users can view the sensitivity profile of a cell line of interest relative to other CRC cells.



catenin (cadherin-associated protein), beta 1, 88kDa

The protein encoded by this gene is part of a complex of proteins that constitute adherens junctions (AJs). AJs are necessary for the creation and maintenance of epithelial cell layers by regulating cell growth and adhesion between cells. The encoded protein also anchors the actin cytoskeleton and may be responsible for transmitting the contact inhibition signal that causes cells to stop dividing once the epithelial sheet is complete. Finally, this protein binds to the product of the APC gene, which is mutated in adenomatous polyposis of the colon. Mutations in this gene are a cause of colorectal cancer (CRC), pilomatrixoma (PTR), medulloblastoma (MDB), and ovarian cancer. Three transcript variants encoding the same protein have been found for this gene.[provided by RefSeq, Oct 2009]

Gene official symbol	CTNNB1
Entrez gene ID	1499
Gene aliases	CTNNB, MRD19, armadillo
Chromosome	3
Map location	3p21
Gene type	protein-coding

Statistics for CTNNB1

Domains	1
PTMs	36
Protein entries	716
Sequence variants	163
Cell lines/tissues	136
Pathways involved in	21

Post-translational modification (PTMs) in CTNNB1

Filter by
PTMs affected by mutations
PTMs not affected by mutations

Show 10 entries

RefSeq ID	Amino acid sequence variants	PTM type	PTM position	Amino acid	Upstream enzy	PubMed Ids
NP_001895	p.S29Y	Phosphorylation	29	S	CSNK2B	PubMed
NP_001895	p.S29Y	Phosphorylation	29	S	CSNK2A1	PubMed
NP_001895	p.S33C	Phosphorylation	33	S	GSK3B	PubMed
NP_001895	p.S33Y	Phosphorylation	33	S	GSK3B	PubMed
NP_001895	p.S33F	Phosphorylation	33	S	GSK3B	PubMed
NP_001895	p.S37P	Phosphorylation	37	S	GSK3B	PubMed

Domains for CTNNB1 proteins

RefSeq ID: NP_001895

List of CTNNB1's sequence variants

Show 10 entries

Search: CW-2

CSV PDF Print

Type of sequence variant	Sequence variant effect	Genomic position	Nucleotide change	Amino acid change	Cell line/tissue	Mutation source
Substitution	Nonsynonymous	41277276	c.1745A>G	p.R582Q	CW-2	Klijn et al., 2015

Showing 1 to 1 of 1 entries

Previous 1 Next

Figure 3-2: PTMs and domains in β-catenin are affected due to mutation. Snapshot of β-catenin molecular page is displayed. The PTMs affected by mutations can be viewed in the tab PTMs. Mutations in β-catenin at positions important for phosphorylation (S33, S37, T41 and S45) allows for the stabilization of β-catenin and constitutive activation of the Wnt signaling pathway. The upstream kinases responsible for the phosphorylation is also provided along with the literature reference. Likewise, mutations in the armadillo domain can be viewed by correlating the sequence variants and the domain span regions. For

example, mutations in the armadillo domain (p.R582) in β-catenin have been described which have been reported to alter the binding of β-catenin to TCF4 [235]

3.7. Accessing Colorectal Cancer Atlas

Users can search Colorectal Cancer Atlas through the home, query or browse pages (Figure 3-3). In addition, the website features a navigation menu and a search box at the top of the page. The database can be queried by gene symbol, Entrez Gene ID, protein name, cell line name or pathway. The browse page provides users with the option to access the database by categorised lists of genes, sequence variations, cell lines and techniques. The browse page allows the users to search for sequence variations in genes of interest and displays them in interactive color-coded table format. The gene information page includes gene details, associated GO terms, sequence variations (displayed in an interactive table), domain details, PTMs, a protein data page leading to experimental techniques and quantitative data with an interactive heat map, a column chart for spectral abundance and a list of detected peptides. Other information includes a list of cell lines and tissues that contain sequence variants in a given gene, a list of pathways in which the gene is involved, and an interactive protein-protein interaction network for the protein encoded by the gene. The cell line page provides details of the cell line, an interactive table of gene sequence variants identified in the cell line, an interactive table of dysregulated pathways and 5-FU drug sensitivity profile. Data curated in Colorectal Cancer Atlas is available as tab-delimited files and is free for download to all users. Using the custom database option, the tab delimited data can also be uploaded into FunRich [240], a functional enrichment analysis tool to identify classes of genes/proteins that are overrepresented in a specific category.

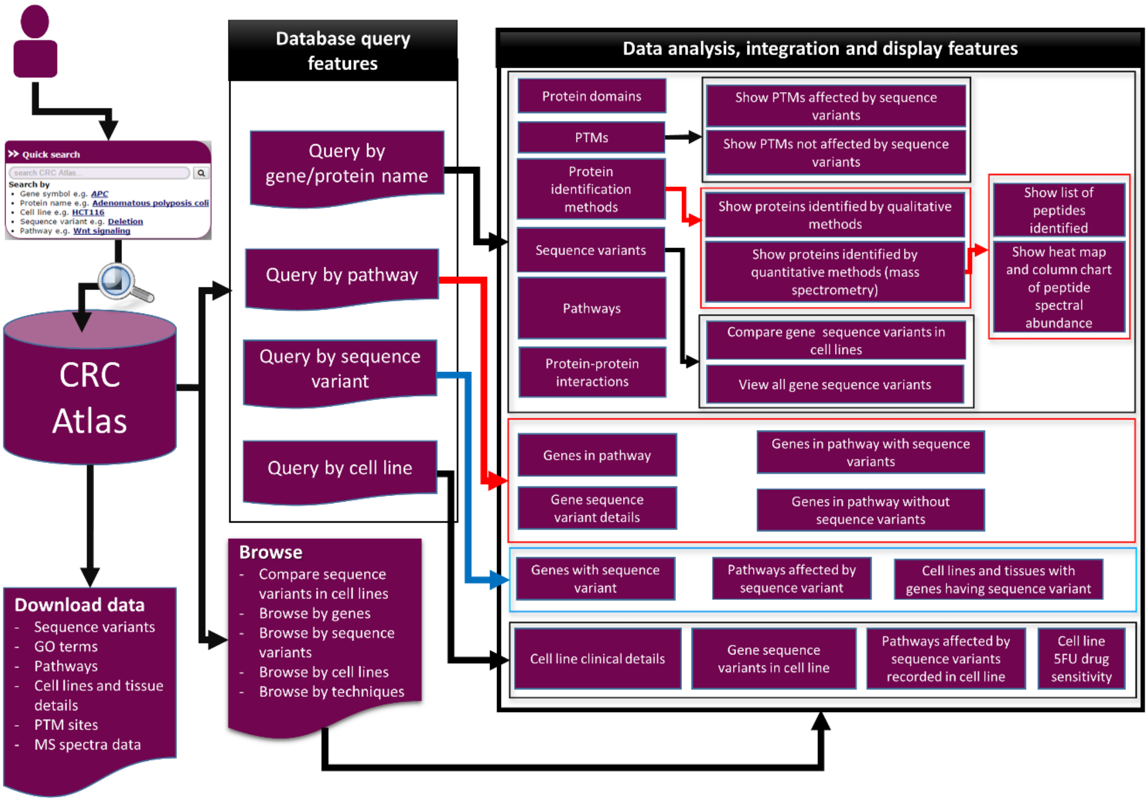

Quick search
search CRC Atlas...
Search by
Gene symbol e.g. APC
Protein name e.g. Adenomatous polyposis coli
Cell line e.g. HCT116
Sequence variant e.g. Deletion
Pathway e.g. Wnt signaling
CRC Atlas
Download data
- Sequence variants
- GO terms
- Pathways
- Cell lines and tissue details
- PTM sites
- MS spectra data
Database query features
Query by gene/protein name
Query by pathway
Query by sequence variant
Query by cell line
Browse
- Compare sequence variants in cell lines
- Browse by genes
- Browse by sequence variants
- Browse by cell lines
- Browse by techniques
Data analysis, integration and display features
Protein domains
PTMs
Protein identification methods
Sequence variants
Pathways
Protein-protein interactions
Show PTMs affected by sequence variants
Show PTMs not affected by sequence variants
Show proteins identified by qualitative methods
Show proteins identified by quantitative methods (mass spectrometry)
Compare gene sequence variants in cell lines
View all gene sequence variants
Show list of peptides identified
Show heat map and column chart of peptide spectral abundance
Genes in pathway
Gene sequence variant details
Genes in pathway with sequence variants
Genes in pathway without sequence variants
Genes with sequence variant
Pathways affected by sequence variant
Cell lines and tissues with genes having sequence variant
Cell line clinical details
Gene sequence variants in cell line
Pathways affected by sequence variants recorded in cell line
Cell line 5FU drug sensitivity

Figure 3-3: Use case for Colorectal Cancer Atlas. Users can access Colorectal Cancer Atlas through the query or browse pages. The browse page provides users with the option to access the database by categorised lists of genes, sequence variations, cell lines and techniques. Further to this, sequence variations in specific genes can be viewed as an interactive table format. The gene information page includes gene details, associated GO terms, sequence variations, domain details, PTMs, a protein data page leading to experimental techniques and quantitative data with an interactive heat map, a column chart for spectral abundance and a list of detected peptides. The entire data in Colorectal Cancer Atlas can be downloaded as tab-delimited files.

3.8. Future directions

Colorectal Cancer Atlas will be continuously updated with more studies as they become available and additional features. Studies currently being curated include Wnt signaling activity determined by the TOPFLASH assay, and genomic and proteomic data generated from patient derived xenografts.

Chapter 4

Perturbation of protein-protein interaction network based on APC mutations in colorectal cancer

This chapter has been submitted for consideration as a journal article to Scientific Reports

David Chisanga[1], Shivakumar Keerthikumar[3], Naveen Chilamkurti[1] and Suresh Mathivanan[2*]

[1]Department of Computer Science and Information Technology, La Trobe University, Bundoora, Victoria, 3086, Australia

[2]Department of Biochemistry and Genetics, La Trobe Institute for Molecular Science, La Trobe University, Melbourne, Victoria, 3086, Australia

[3]Computational Cancer Biology Program, Cancer Research Division, Peter MacCallum Cancer Centre, Melbourne, VIC 3000, Australia

*To whom correspondence should be addressed

Dr Suresh Mathivanan

Department of Biochemistry and Genetics,

La Trobe Institute for Molecular Science,

La Trobe University,

Bundoora, Victoria 3086, Australia

Tel: +61 03 9479 2565

Email: S.Mathivanan@latrobe.edu.au

4.1. Abstract

Colorectal cancer (CRC) is the third-most common form of cancer in the world with a high rate of morbidity and mortality. The majority (95%) of CRCs are adenocarcinomas whose pathogenesis is preceded by growths in the linings of the bowel called polyps caused by either inherited or somatic genetic alterations. Among the several genes implicated in CRC, mutations in the APC gene have been shown to be the precursor to the cascade of changes that the polyps undergo. Furthermore, research has shown that adenomatous polyposis coli (APC) does not act alone but rather regulates other genes such as β-catenin (CTNNB1) of the Wnt signaling pathway, a pathway that is responsible for regulating cellular behaviours such as cell migration, cell polarity, and organogenesis. The dysregulation of the Wnt signaling pathway has been implicated in CRC.

Interestingly, significant progress has been made in characterising the roles of APC in CRC, nonetheless, there are ongoing efforts to better understand the mechanisms of gene networks behind the proliferation and viability of tumours in CRC. Since cancer is known as the disease of the pathways, in this study, we developed a novel network analysis method that integrates genomics and proteomics data to analyse the topological changes in a PPI network when APC is mutated. The aim was to identify genes that are essential for the proliferation and viability of tumours in CRC. Using this method, we identified new and already known genes which are essential for the proliferation of CRC. We also identified pathways that are significantly affected by the topological changes induced by the mutation of APC. The roles of the predicted genes in the proliferation and viability of tumours in CRC were validated using the Achilles dataset. Upon validation, over 10 unique genes were shortlisted as being essential in the viability and proliferation of tumours in CRC. Notable among these included DKK3, KRT23, STAT3, TSG101, APOBEC3G and ASL.

4.2. Introduction

Colorectal cancer (CRC), also known as bowel cancer is the third-most common form of cancer in the world and has one of the highest rates of cancer related morbidity and mortality around the world [3]. For instance, in 2015 alone, there were more than 774,000 colorectal cancer-related deaths in the world, making it the third-leading cause of cancer deaths that year after lung and liver cancers.

According to the World Cancer Research Fund International, about 95% of CRCs are adenocarcinomas, and their pathogenesis starts with growths in the linings of the bowel called polyps, which result from either inherited or somatic genetic alterations [241]. The polyps gain additional alterations, from being adenomas, they develop into adenocarcinomas and ultimately become metastatic. Genes considered to be the key drivers and have been found mutated in CRC include: APC, TP53, KRAS, PIK3CA, FBXW7, SMAD4, TCF7L2 and NRAS. Other genes like CTNNB1 (β-catenin), SMAD2, FAM123B and SOX9 have also been found to be mutated and have been implicated in CRC [28]. However, one of the significant challenges in cancer medicine has been to understand how these driver genes function in physiological states and how this function is disrupted in pathological states. Among the key CRC driver genes, APC has been found to have one of the highest frequency rates of mutation in >80% of sporadic CRCs [25] and mutations in APC (which lead to the loss of its functionality) are considered to be the precursor to the cascade of changes that the polyps undergo [17]

The characterisation of the role of APC in CRC and other forms of cancer has been well documented in the literature [17, 18, 25]. It has been shown that APC does not act alone but rather by regulating other genes such as CTNNB1 and hence the Wnt signaling pathway, a pathway that is responsible for regulating cellular behaviours such as cell migration, cell polarity, and organogenesis. APC therefore indirectly controls the Wnt pathway through its regulation of CTNNB1 and thus regulates functions such as cell adhesion and migration, and signal transduction as well as other additional functions like microtubule assembly and chromosome segregation. Consequently, when APC becomes mutated, it loses its functionality to regulate the different range of functions that it regulates, top among these is the loss of functionality to regulate the Wnt pathway which is associated with CRC [35, 36].

While significant progress has been made in characterising essential genes in CRC as researchers continue to perform global analyses of various omics related data, the general characterisation of genes as biomarkers is however hindered by the heterogeneous nature of cancer whereby individuals have different forms and stages of the same disease. Because cancer is a disease of the pathways [8], given the fact that genes are part of a nonlinear set of interconnected pathways and perform their functions through a complex cellular network, there is a need to better understand the underlying mechanism of gene networks in both physiological and pathological states.

In this study, we build on the work done in Chapter 3 and use protein-protein interactions (PPI) to study the dynamic network changes that take place in protein-protein interactions when APC is mutated. The aim is to perturb the PPI network with APC mutational information and understand how the PPI network topological structure changes when APC is mutated. PPIs are a result of two or more proteins binding together purposely to carry out a specific biological function in a cell [41]. Given the enormous volumes of heterogeneous data that is continuously being churned out of research laboratories around the world to understand cancer, PPI networks provide us with an opportunity to integrate the various forms of data and form a pictorial representation of cellular function and other biological process changes.

4.3. Results

4.3.1. Profiling APC as a driver gene in colorectal cancer

Mutation frequencies of the genes APC, FBXW7, KRAS, NRAS, PIK3CA, SOX9, SMAD4, TCF7L2, TP53 and FAM123B (which are driver genes in colorectal cancer) were profiled in over 600 TCGA colorectal cancer samples. For each sample, the occurrence of a gene mutation was counted as one, regardless of the number of mutation occurrences. Based on these results, APC, TP53 and KRAS were found to have high rates of mutation among the TCGA samples with each gene having mutation frequencies of 27%, 21% and 15% respectively as summarised in Figure 4-1 (a). Based on these observations as well as previous observations [30, 32] where it has been shown that APC is one of the most frequently mutated genes in CRC patients, TCGA samples with mutant APC were selected for further analysis. PPI networks and network theory methods were used to analyse the topological changes that take place in APC mutant samples.

To begin, we first performed a pathway enrichment analysis using FunRich [242] for APC and its interacting partners to understand the biological processes and pathways that they are involved in physiological conditions. Pathways enriched among the APC interacting partners included: the destruction complex, Wnt signalling network, E-cadherin signalling, and Syndecan-4-mediated signalling. Using these pathways, we clustered APC interacting partners into groups based on the pathways in which they are involved, as shown in Figure 4-1 (b). In addition, using the RNA-seq expression data from the SW480+APC cell line, we overlayed it over the APC subnetwork established in the materials and methods section to understand the expression profile of APC interacting partners when APC functionality is restored, as shown in Figure 4-1 (c). The expression profile of the APC subnetwork in Figure 4-1 (c) was taken to be the standard normal APC subnetwork if APC were functioning normally.

We further performed differential gene expression analysis between the SW480 + APC restored and SW480 cell line with mutant APC, and genes which had an absolute fold change >=2 and had a p-value<0.05 were marked as being differentially expressed. Of the over 12,500 genes, 738 were found to be overexpressed while 957 were found to be underexpressed. The differential gene expression results were then used for further downstream analysis, as described in the next sections.

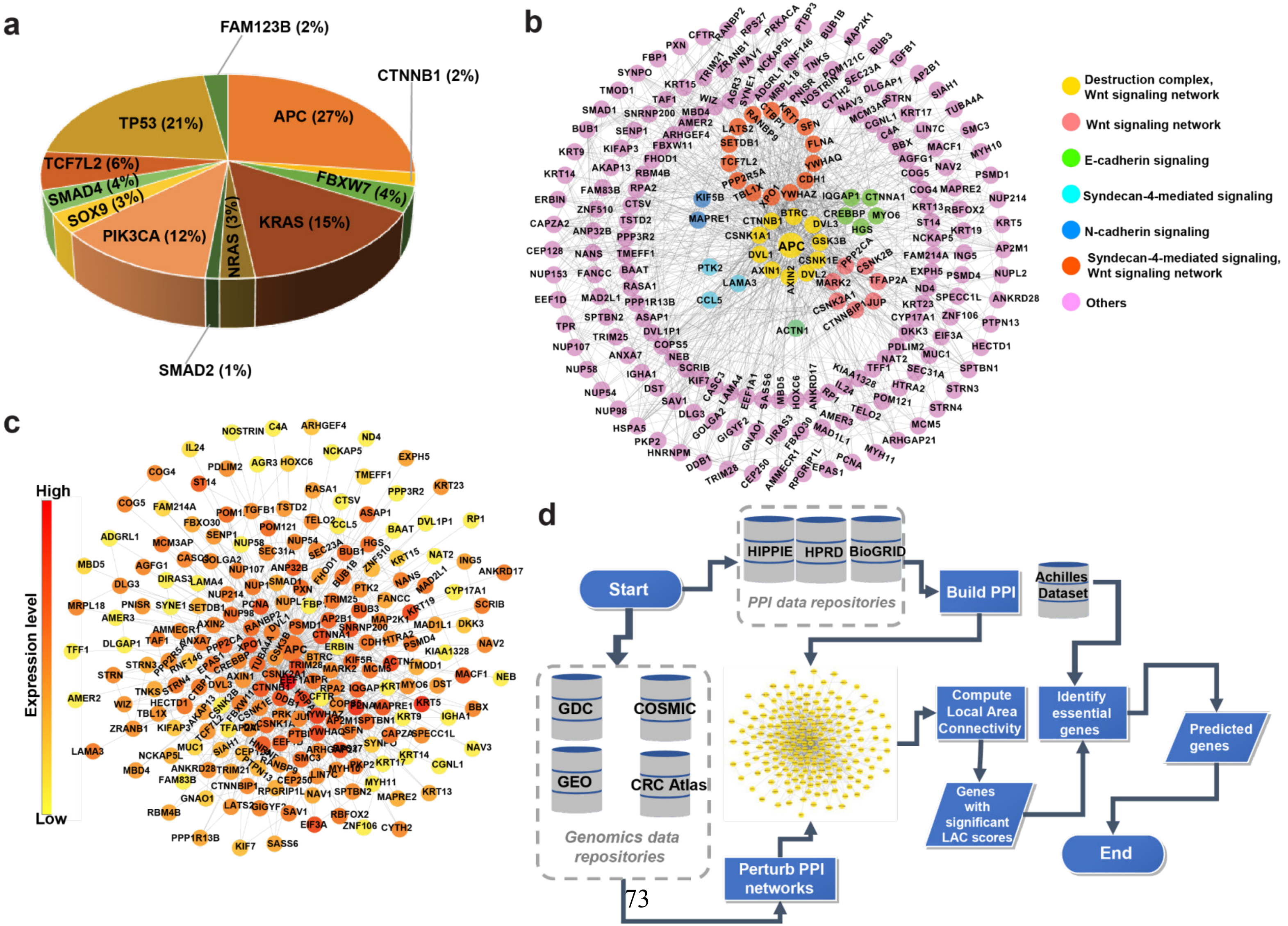
a
FAM123B (2%)
CTNNB1 (2%)
APC (27%)
TP53 (21%)
TCF7L2 (6%)
SMAD4 (4%)
SOX9 (3%)
FBXW7 (4%)
PIK3CA (12%)
NRAS (3%)
KRAS (15%)
SMAD2 (1%)
b
APC
Destruction complex, Wnt signaling network
Wnt signaling network
E-cadherin signaling
Syndecan-4-mediated signaling
N-cadherin signaling
Syndecan-4-mediated signaling, Wnt signaling network
Others
c
High
Expression level
Low
d
Start
HIPPIE
HPRD
BioGRID
PPI data repositories
Build PPI
Achilles Dataset
GDC
COSMIC
GEO
CRC Atlas
Genomics data repositories
Compute Local Area Connectivity
Identify essential genes
Predicted genes
Genes with significant LAC scores
Perturb PPI networks
End

Figure 4-1: Characterisation of the APC subnetwork (a) The mutation frequency of commonly mutated genes in CRC was profiled in CRC TCGA patient samples. APC had one of the highest rates of frequency of mutation followed by TP53 and KRAS. (b) The direct interacting partners of APC and their interactions were obtained, and pathway enrichment analysis performed. An APC subnetwork was generated and proteins clustered by the enriched pathways. (c) Gene expressions of all APC interacting partners in the SW480+APC were obtained and overlayed over the APC subnetwork with the colour intensity depicting genes that were highly and lowly expressed. (d) Summary of the workflow that was followed in identifying essential genes. Genomics data for SW480 cell lines were downloaded from GEO and differential gene expression analysis was performed between SW480+APC and the SW480 with the defective APC. TCGA CRC patient data and cell line data were also downloaded from GDC, COSMIC and CRC atlas. PPI data was also downloaded from three data repositories which were then used to build a weighted PPI. The weighted PPI was then perturbed with APC mutation information and differential gene expressions with LAC being used to compute topological changes. Genes with significant topological changes (LAC score changes) were then validated against the Achilles dataset to identify genes which may be essential to the viability of CRC when APC is mutated.

4.3.2. Perturbation of PPI networks in APC mutant samples

Local area connectivity (LAC) was computed as summarised in Figure 4-1 (d) and described in the materials and methods section to quantify the global topological changes in a protein-protein interaction network when APC is mutated. Weighted PPI networks were created for each TCGA patient sample and cell line sample, and overlayed with gene mutation information, the differential gene expression status as well as cancer gene census information. The same procedure was also repeated for the SW480+APC cell line. Using the PPIs, LAC scores were calculated for mutant APC TCGA samples and CRC cell lines as well as for the SW480+APC cell line.

We then measured the variability in the LAC scores from the TCGA samples and cell lines against that of the SW480+APC cell line by calculating the z-scores as outlined in the methods section. Genes with a z-score >=2 or <=-2 were considered to be significant and were selected for further downstream analysis. From the over 16,000 genes included in the PPI networks, 1,837 genes were found to have absolute z-scores >=2 in the TCGA samples analysis while 2,289 genes had absolute z-scores >=2 in the cell lines. When filtered for common genes between the two result sets, 1,649 genes remained out of which 965 had negative z-scores, and 684 had positive z-scores, as shown in Figure 4-2 (a) and supplementary table 4.1.

Next, we performed a series of downstream analyses to validate and characterise the roles of these genes in CRC, as discussed in the next sections.

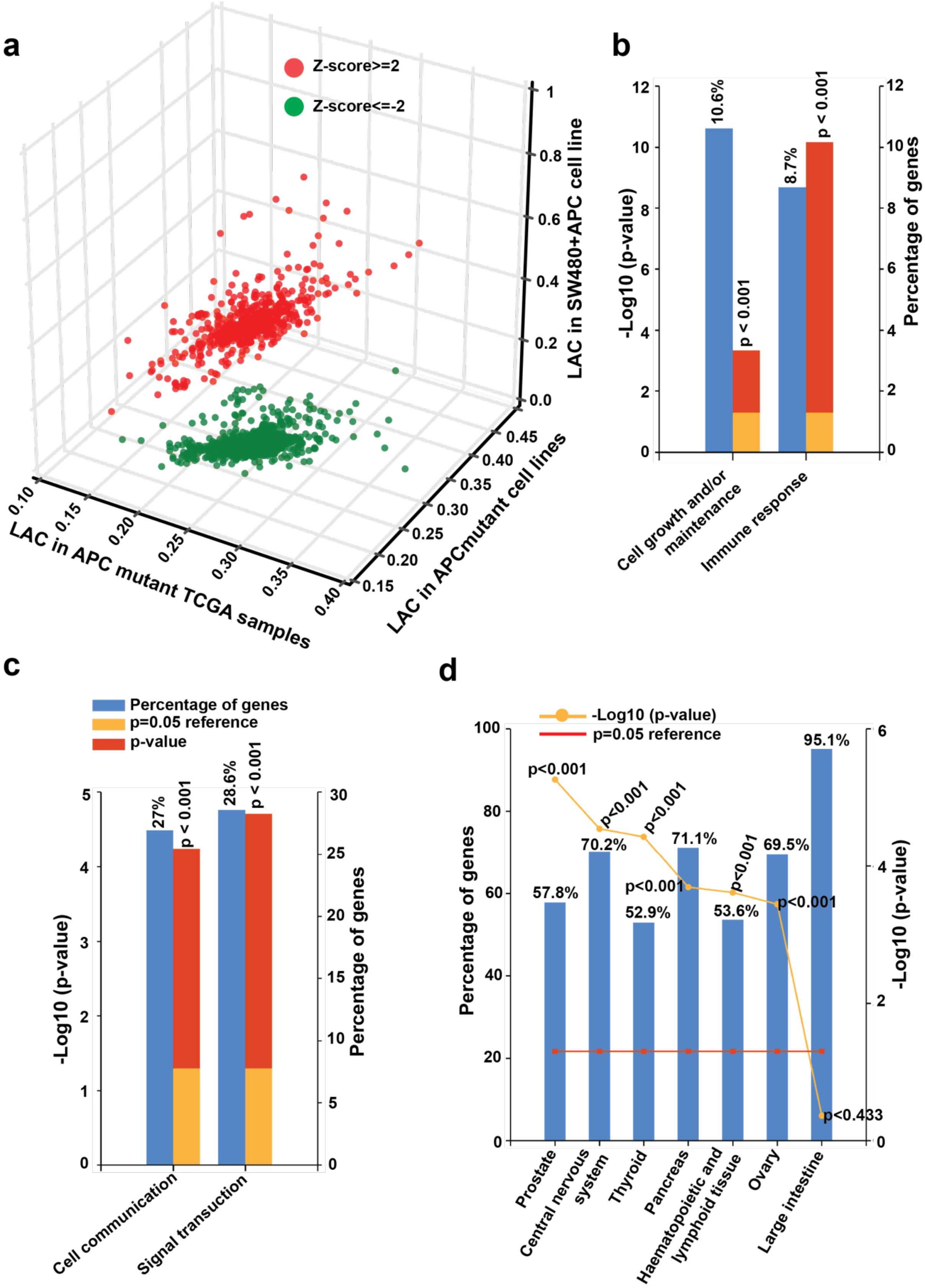


Figure 4-2: Perturbation of PPI network with APC mutational information. (a) Computed LAC scores in TCGA and cell line samples are compared against the LAC scores in

SW480+APC cell line using z-score. Genes with significant changes in LAC scores are then split between those with z-scores>=2 (in red) and those with z-scores<=-2 (in green). (b) Using FunRich, we performed pathway enrichment analysis for the genes with positive z-scores. These were significantly enriched for cell growth and/or maintenance and immune response. (c). Using FunRich, we also performed pathway enrichment analysis for genes with negative z-scores. These were significantly enriched for cell communication and signal transduction (d). COSMIC enrichment of the identified genes in (a) showed a number were significantly enriched in various forms of cancer.

4.3.3. Pathway enrichment

To further characterise the genes identified above, we performed a pathway enrichment analysis to understand the pathways that are affected by the topological changes in the PPI network because of APC mutations. To do this, we split the genes into two groups: the first group consisted of genes that had positive z-scores, that is, genes whose average LAC scores in both TCGA samples and cell lines was less than that in the SW480+APC cell line; the second group consisted of genes that had negative z-scores implying that these genes had average LAC scores in both TCGA samples and cell lines that were greater than those in the SW480+APC cell line. For the first group, two pathways were found to be significantly enriched ($p<0.05$) and comprised cell growth and immune response, as shown in Figure 4-2 (b); while in the second group, the pathways found to be significantly enriched ($p<0.05$) included cell communication and signal transduction, as shown in Figure 4-2 (c).

In addition, we also performed enrichment analysis of the COSMIC cancer gene census list to determine how many of the identified genes were implicated in cancer. The enrichment showed that 55 genes were part of the cancer gene census list as shown in supplementary table 4.1 while several of the other genes were also implicated in various cancer types, as shown in the enrichment bar graph in Figure 4-2 (d).

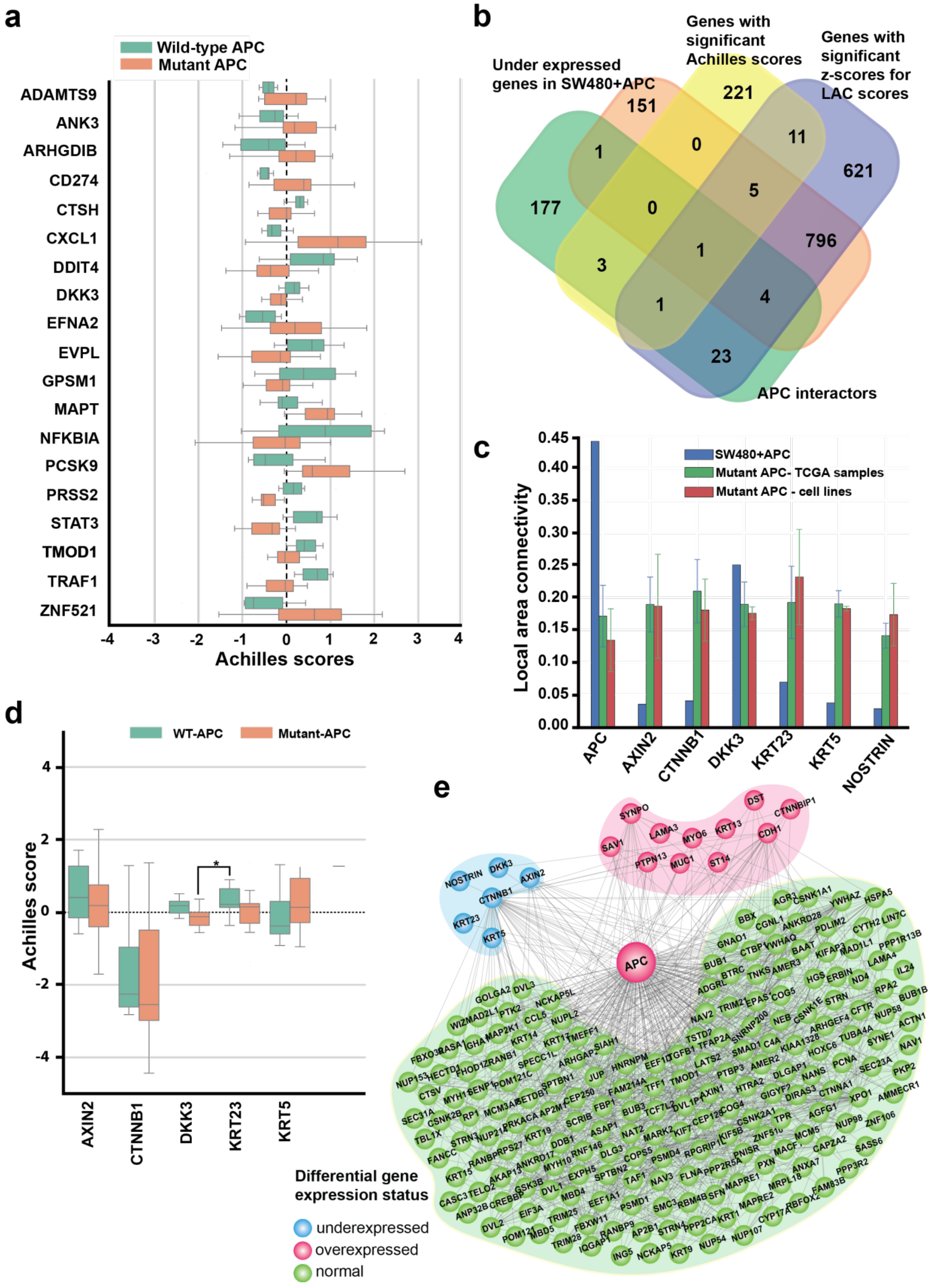


Figure 4-3: Identifying genes essential for cell survival in mutant APC cell lines. (a) Using data from Project Achilles (https://portals.broadinstitute.org/achilles), we compared (using t-test) the essentiality of the genes with significant LAC scores to the viability of cells in

APC mutant cell lines against wild type APC cell lines. The genes in the boxplot were found to have significant differences in the Achilles score between mutant APC cell lines and wild-type APC cell lines. Genes were the median score in mutant APC cell lines was less than that in mutant-APC cell lines were selected for further analysis (b) Next, we integrated the various result sets to identify genes which were significantly enriched in all/or some of the result sets. Genes which were found to be common in any three of the datasets were selected for further analysis. (c). The LAC for APC interactors which were found to be differentially underexpressed were compared across the 3 data sets with AXIN2, CTNNB1, KRT23, KRT5 and KRT23 all showing an increase in LAC when compared to that in SW480+APC while DKK3 showed a decrease in connectivity. (d) APC interactors which were differentially expressed were selected and their Achilles scores compared. Only genes which were underexpressed in the SW480+APC cell line were tested for essentiality in cell viability. DKK3 was found to have a significant difference between Achilles scores in mutant APC cell lines and wild-type APC cell lines. (e) APC subnetwork is overlayed with differential gene expression and genes are then clustered based on the differential gene expression status. Genes in clustered with the blue colour were found to be underexpressed while those clustered with the pink colour were found to be overexpressed in the SW480+APC cell line .

4.3.4. Identifying genes essential for cell survival in mutant APC cell lines

To understand the role of the genes selected from the above in CRC, we analysed their effect on tumour viability and proliferation using the Achilles dataset which was discussed in the materials section. From the Achilles dataset, we obtained all known CRC cell lines and split them into two groups, consisting of mutant APC cell lines and wild-type APC cell lines, respectively. For each gene with absolute z-scores >=2, we compared the Achilles scores between APC mutant cell lines and APC wild-type cells. 19 genes were found to have significant differences ($p<0.05$) between Achilles scores in APC mutant cell lines and APC wild-type cell lines, namely DKK3, MAPT, ZNF521, GPSM1, CTSH, EFNA2, EVPL, ANK3, RHGDIB, CD274, STAT3, NFKBIA, CXCL1, TRAF1, PRSS2, DDIT4, TMOD1, PCSK9, and ADAMTS9 as depicted in the boxplots in Figure 4-3 (a). From these, genes that had a significantly lower median Achilles score in the mutant APC cell lines than that in wild-type APC cell lines were selected, these being: DKK3, GPSM1, CTSH, EVPL, STAT3, NFKBIA, TRAF1, PRSS2, DDIT4, and TMOD1. The lower Achilles scores for these genes in mutant cell lines implies that when knocked-out or knocked-down in their respective mutant APC cell lines, the rate of cell proliferation in mutant cell lines is less than that in wild-type APC cell lines.

Additionally, we also probed the genes identified using LAC for APC interactors to confirm how many had significant changes in their LAC scores. Of the 205 APC interactors in the PPI network, 38 were found to have significant changes in the LAC scores. These were then compared against APC interactors that were differentially expressed in the SW480+APC cell line. Of the 18 differentially expressed APC interactor genes, 14 were found to be shared between the two result sets, as shown in Table 4-1.

For this work, we focused on genes that were found to be underexpressed in the SW480+APC cell line and had variable LAC scores, as summarised by the Venn diagram in Figure 4-3 (b), including DKK3, KRT23, KRT5, CTNNB1 and NOSTRIN is shown in Figure 4-3 (c). For each gene, we compared its Achilles scores between the two groups, as shown by the boxplots in Figure 4-3 (d). The Achilles median score for DKK3 in mutant APC cell lines was significantly ($p<0.05$) less than that in wild-type APC cell lines while the median scores for CTNNB1 in both groups were significantly lower than any of the selected genes. However, there was no record of NOSTRIN in the Achilles dataset and KRT5's median score in APC mutant cell lines was higher than that in APC wild-type cell lines, implying that it does not affect cell viability when APC is mutated. These results, therefore, indicate that when DKK3 is knocked-out in APC mutant cell lines, tumour viability and progression is reduced in several of the mutant APC cell lines, which means that DKK3 is essential to the viability and proliferation of cells in CRC cell lines. The scores for CTNNB1 on the other hand which were significantly low in both groups showed that it is essential to the viability of cells in both groups. KRT23, on the other hand, showed a small variation in the distribution of scores across the cell lines in the two groups. Many of the mutant APC cell lines had negative scores when compared to those with wild-type APC leading us to conclude that KRT23 may also be essential to the viability of mutant APC cell lines. Figure 4-3 (e) provides a summary of APC interactors and their differential gene expression status when APC functionality is restored.

Table 4-1: Differentially expressed APC interactors. Genes which had significant local area connectivity changes were differentially expressed in wild-type APC cell lines

(SW480+APC) and were checked against the Achilles dataset for their effect on cell viability in mutant APC cell lines.

Entrez ID	Symbol	Log FC	LAC-SW480	LAC-TCGA	LAC-Cell lines
56998	CTNNBIP1	2.32	0.42	0.22	0.20
60485	SAV1	1.19	0.40	0.19	0.19
3860	KRT13	2.78	0.41	0.19	0.18
6768	ST14	3.76	0.46	0.19	0.26
25984	KRT23	-3.64	0.07	0.19	0.23
3852	KRT5	-1.67	0.04	0.19	0.18
1499	CTNNB1	-1.65	0.04	0.21	0.18
5783	PTPN13	3.94	0.44	0.20	0.18
3909	LAMA3	2.28	0.50	0.18	0.18
4646	MYO6	1.48	0.44	0.20	0.19
11346	SYNPO	1.90	0.48	0.22	0.22
115677	NOSTRIN	-1.69	0.03	0.14	0.17
4582	MUC1	2.52	0.47	0.22	0.20
27122	DKK3	-1.21	0.25	0.19	0.18

We further characterised the genes found to be under-expressed in the SW480+APC cell line but are not direct interacting partners of APC against the list of genes with variable LAC scores. As above, we compared their scores in mutant APC CRC cell lines against their scores in wild-type APC cell lines using independent samples t-test. Of the 951 under-expressed non-APC interactors, GPMS1 a second order interacting partner of APC was found to have a significant difference in the Achilles scores between APC mutant and APC wild-type. As such, we also compared the Achilles scores for the second order interacting partners of APC, from which 50-second order APC interacting partners were found to have significant differences in the Achilles scores between APC mutant and APC wild-type cell lines, as shown by the boxplots in Figure 4-4 (a). A literature search of the 50 genes showed that 16 (APOBEC3G, ASL, CDKN1A, DYNC1H1, ELMO3, FUBP1, GNA11, NFKBIA, PSMA1, STAT3, SUPT5H, TRAF1, TRIP13, TSG101, TUBA1B and USP39) are implicated in colorectal cancer.

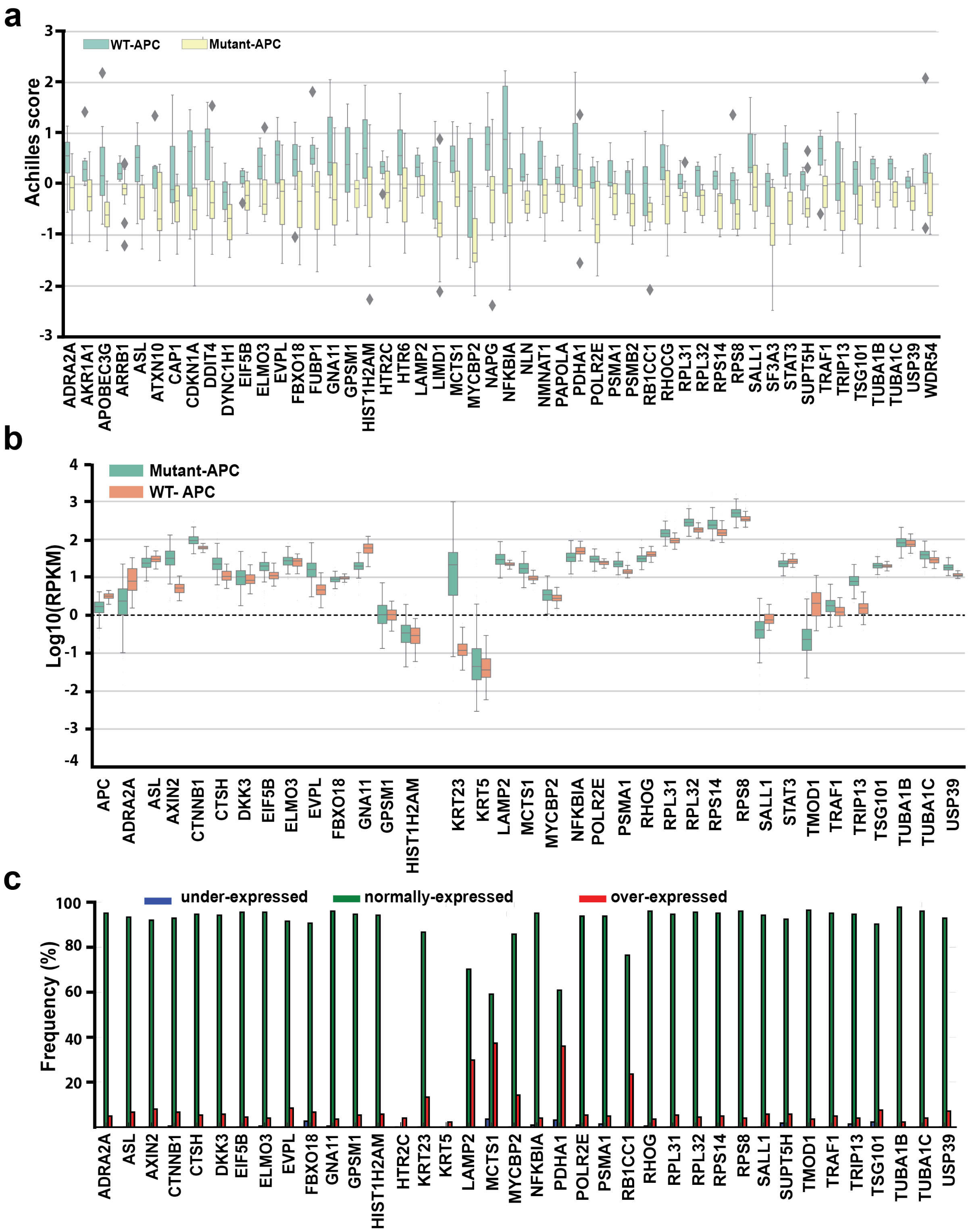


Figure 4-4: Genomic profiling of identified genes in TCGA samples Perturbation of PPI networks in APC mutant samples. (a) Gene essentiality was also performed for second-order APC interacting partners. Genes with significant Achilles scores in mutant APC cell lines and wild-type cell lines were selected as potential candidates as well (b) All the genes that were selected as potential candidates were then profiled in TCGA samples for their

frequency of being differentially expressed. (c) The expression profile of the selected genes from (b) is compared between mutant-APC samples and wild-type APC samples.

4.3.5. General profiling of selected genes in TCGA patients with APC mutations

To further characterise the shortlisted genes, we profiled their gene expression in TCGA CRC patient samples as discussed in the materials section. We compared their gene expressions in TCGA samples with wild-type APC samples against those with mutant APC, as shown by the boxplots in Figure 4-4 (b). The results showed that some of the genes had higher median expression values in mutant APC TCGA samples when compared to wild-type APC samples, including AXIN2, CTNNB1, CTSH, DKK3, EIF5B, EVPL, KRT23, MCTS1, PSMA1, RPL32, RPS14, RPS8, and TRIP13.

Furthermore, we also wanted to find out how frequently these genes were differentially expressed across the TCGA samples. We analysed their differential expression status among TCGA samples, as summarised in Figure 4-4 (c). It was found that while all the genes were normally expressed in many of the TCGA samples analysed, genes with high median expression values in mutant APC samples were found to be over-expressed in a number of mutant samples which, therefore, implies that when APC is mutated, the selected genes are either normally or overexpressed which is consistent with the observations in the previous section, where it is shown that when APC functionality is restored, these genes are downregulated when compared to the mutant cell line.

4.4. Discussion

The identification of essential genes in the tumorigenesis, proliferation and metastasis of cancer remain one of the significant challenges in cancer research due to the heterogeneous nature of cancer. Mutations in APC have been shown to be essential in the tumorigenesis of CRC [243]. It is against this background that in this study, we set out to understand the topological changes that take place in PPI networks when APC is mutated and therefore, attempted to identify genes that are essential for the proliferation and viability of tumours in CRC cells when APC is mutated. With the rapid increase and availability of heterogeneous omics datasets deposited in various online repositories, techniques to reuse and integrate such types of data to gain insights from them have become more than necessary.

In this work, we used the concept of the ‘small-world’ [244] property of PPI networks, whereby mutational changes in one gene are cascaded beyond its direct interacting partners. We, therefore, sought to define a method that could integrate the various omics data while at the same time quantifying the dynamic changes that take place in a PPI network due to mutational and differential gene expression changes. Here, we have developed a novel method, local area connectivity (LAC), that perturbs a PPI network by globally cascading APC’s mutation information through a PPI network and quantifies the topological changes arising from the perturbation. The method uses node degree to calculate the connectivity of a node but in the process, penalises interacting partners that are either downregulated because of a mutation in APC or are known to be cancer related and are either themselves mutated or not.

Using this method, we predicted over 1600 genes as having significant topological changes in their local connectivity which included already known candidates as well as new potential candidates. Enrichment analysis of the predicted results showed that they were significantly enriched for such biological pathways as cell growth, immune response, cell communication and signal transduction which are all considered to be important in the tumorigenesis, proliferation and metastasis of cancer [16]. We found that genes whose high average LAC samples were highly enriched for signal transduction and cell communication while those with low LAC scores were highly enriched for immune response and cell growth/maintenance. Here, we hypothesised that when APC is mutated, genes which are involved in the signal transduction and cell communication pathways become highly interconnected as a result of APC mutations and consequently, become irresponsive to new signals as the signalling processes are never terminated. This view is supported by several previous research studies which documented the roles of these pathways in cancer [18, 245]. On the contrary, when APC is mutated, genes which are involved in immune response and cell growth/maintenance are less well connected, and as such, their respective pathways may become inactivated which, in turn, may enhance the proliferation of cancer cells as the mechanisms responsible for controlling cell growth and/or maintenance or immune response are lost [246]. However, for the immune response, another explanatory reason as to why it is highly enriched in genes with low average LAC scores maybe due to the fact gene expression profiling was conducted in a cancer cell line with APC functionality restored.

To validate the predicted results, we used the Achilles dataset and found several of the predicted genes to be essential in the proliferation of CRC cancer cell lines. Among the identified genes included direct APC interactors such as AXIN2, CTNNB1, DKK3, KRT23, KRT5 and NOSTRIN which were also found to be underexpressed in the SW480+APC cell line. The roles of AXIN2 and CTNNB1 in CRC have been well documented in the literature [33, 243, 247, 248]. We, therefore, used these as references to understand the roles of the other genes when analysing the Achilles data set and as a result, we found that DKK3 and KRT23 are essential in cell proliferation in mutant APC cell lines. This observation was confirmed further by results from the gene expression profiling of TCGA patient samples which showed that the median expressions of DKK3 and KRT23 in mutant APC samples are higher than that in wild-type APC samples. This observation is supported by previous research [249] where it has been shown that DKK3 is overexpressed in CRC and Birkenkamp-Demtröder, et al. [250] also showed that by knocking down KRT23 in CRC samples, cell proliferation is reduced. NOSTRIN and KRT5 did not show notable differences in their Achilles scores.

In addition to APC interactors, we also profiled genes that are not direct interactors of APC but had significant LAC scores and were also found to be essential for the viability of CRC cancer cell lines, notable among these included: STAT3 [251] and TSG101 [252] which have been implicated in CRC. Others included: APOBEC3G, ASL, CDKN1A, CTSH, DDIT4, DYNC1H1, ELMO3, EVPL, FUBP1, GNA11, GPSM1, NFKBIA, PRSS2, PSMA1, SUPT5H, TMOD1, TRAF1, TRIP13, TUBA1B and USP39. A search on PubMed for these genes revealed that they had been found to play a role in CRC or some other form of cancer. For instance, APOBEC3G, a gene that codes for the protein apolipoprotein B mRNA-editing enzyme catalytic polypeptide-like 3G has been found to be associated with poor prognosis in CRC patients [253], ASL a gene that codes for arginosuccinate lyase has also been shown to dysregulate nitric oxide (an essential mediator in the tumorigenesis of various cancers) which, in turn, inhibits the proliferation of cancers such as liver and breast cancer [254]. The remaining genes have also been implicated in functions ranging from tumorigenesis to metastasis [255-259]. While the roles of some of the predicted genes have been documented, there is still a need to understand their roles in the progression of CRC when APC is mutated. In this work, we have shown that when APC is mutated, these genes undergo significant changes to their interactions within a PPI network and are therefore likely to play an essential role in CRC.

In conclusion, while our method successfully identifies genes that are perturbed by mutations in APC, it should be noted that due to the none-availability of the effect of APC mutations on the cell phenotype, all nonsynonymous mutations in APC in this study are treated equally. Furthermore, the protein-protein interactions used are inferred from various cell lines and contexts, and as such, some of the interactions may not take place. In this study, we assume that such interactions do take place in the cell line of interest before overlaying with gene expression and mutation status in the inference of the effect on protein-protein interactions. Caution should therefore be taken when applying local area connectivity to specific studies were the phenotypic effect of a mutation is known or the context of the protein-protein interactions in a given cell line is known.

4.5. Materials and methods

4.5.1. Description of data

Gene expression dataset

Read counts for SW480 cell line RNA-seq data were downloaded from the Gene Expression Omnibus (GEO) database deposited there by King, et al. [260] with the accession number GSE76307. The dataset consists of three samples, each sample having three replicates. The three samples are as follows: SW480 + APC which has APC functionality restored by overexpressing wild-type APC in the SW480 cell line with mutant APC; APC mutant SW480 cell line with the defective APC; and a control vector of the SW480 cell line.

Differential gene expression analysis of SW480 dataset

We performed differential gene expression for the SW480 datasets from the read counts obtained above using edgeR [261], a Bioconductor package in R for performing differential expression analysis. We used the workflow described by Chen, et al. [262] together with the parameters used in [260] which included a selection of genes with at least >1 read per million in a sample as being expressed. We filtered for genes with at least one read per million which resulted in a list of slightly over 12,500 genes from an initial list of over 20,000. We then used the GLM approach in edgeR to calculate differential gene expression between the two groups by TNM normalisation. In our case, we compared gene expressions in APC restored SW480 cell line against those in the defective APC SW480 cell line to get the differential gene expressions. We further performed FDR adjustments to account for multiple testing such that genes that had a p-value <0.05 and a fold-change >=2 were considered to be differentially expressed.

COSMIC dataset

In addition, we also downloaded differential gene expression status and mutation data for cell lines as well as TCGA [28] patient data from an online database of cancer-related data called Catalogue of Somatic Mutations in Cancer (COSMIC version 80) [263]. We then used a list of known colorectal cancer cell lines which we previously catalogued in the colorectal cancer atlas [11] to filter the COSMIC dataset for colorectal cancer-related cell lines only. From this process, we obtained differential gene expression status data for 37 cell lines and 603 patient samples.

Gene mutations landscape dataset

In addition to the gene expression from the COSMIC dataset, we obtained gene mutation data for cell lines and patients with APC mutations. All silent mutations were filtered out, and the binary numbers 1 and 0 were then used to represent the mutation status of genes as either mutated or not mutated, respectively, in a matrix with columns representing the samples while rows represent genes. This data is then used to generate a mutation landscape of genes in cell lines and patients where APC is mutated.

Gene essentiality dataset

We downloaded genomically characterised data from Project Achilles version 2.4 (https://portals.broadinstitute.org/achilles [12]), an online repository from the Broad Institute. The data characterises genes that are essential for the proliferation and viability of cancer cell lines. Project Achilles aims at identifying and cataloguing genes which are essential for the proliferation and viability of cancer. This is achieved by using genome-scale RNAi and CRISPR-Cas9 techniques to knockout or silence individual genes to identify those genes that influence cell survival. Each gene is then scored to signify its effect on cell viability and the lower the score, the higher its effect on cell viability and vice-versa. We filtered the data to include only those cell lines that are known to have mutations in the APC gene. There are other similar studies such as those by Hart, et al. [264], [265] and [266] which also applied CRISPR to the identification of essential genes.

Clinical dataset

We downloaded clinical data for the aforementioned TCGA patients from the Genome Data Commons (GDC) portal (https://portal.gdc.cancer.gov/) and collected clinical information about the stage of each patients' colorectal cancer as well as the status of each patient at the last follow-up.

Protein-protein interaction dataset

Weighted protein-protein interactions were downloaded from Human Integrated Protein-Protein Interaction rEference (HIPPIE version 2.0) [267], an online database repository of weighted protein-protein interactions. The weights between proteins indicate the confidence or the probability of the interactions being reliable. Here, we filtered all interactions which had scores of 0.

4.5.2. PPI network construction

We represented protein-protein interactions as an undirected network G (V, E) where V is the set of proteins and E is the set of edges representing interactions between the proteins. We used Python, a scripting language and Networkx (a package in Python for network manipulation and analysis) to build a network 'G' with the weights between interactions included as edge weights for further analysis. The weighted network G was then normalised using Laplacian normalisation by first converting the network into a Laplacian matrix L, as shown in equation (4.1):

$$L_m = M_D - M_A \tag{4.1}$$

where M_A is the adjacency matrix of G, rows and columns represent the nodes and the interaction between proteins are indicated by either a 0 (if absent) or the aforementioned edge weights (if present). M_D, on the other hand, is the diagonal matrix of M_A. Laplacian normalisation is then performed using the formula shown in equation (4.2):

$$L = M_D^{-1/2} L_m M_D^{-1/2} \tag{4.2}$$

The network G was updated with the new normalised edge weights.

For the SW480 dataset, we classified the gene differential expression into three classes;1, 0 and -1. Genes which had a fold value >=2 and therefore considered to be upregulated are classified as 1 while genes which had a fold value <=-2 are classified as -1 otherwise they are classified as 0. A network G_{sw480} was then generated from the aforementioned network G and each protein in the network labelled per the corresponding class as above.

We then repeated the aforementioned procedure for each of the TCGA patient samples and the cell lines obtained from COSMIC, and the networks generated were added to a new set, N. The G_{sw480} network was taken as the gold standard to measure the dynamic changes in the networks in set N.

4.5.3. Analysis of the APC subnetwork

To understand the differential expression changes among APC interacting partners when APC functionality is restored in the SW480 cell line, we generated a subnetwork of APC and its interacting partners and overlayed differential expression status of genes corresponding to the proteins. Using this network, all proteins whose corresponding gene expression status was underexpressed were obtained and used for further downstream analysis.

Local area connectivity

To understand the topological changes that take place in the networks because of APC mutations, we defined a new metric called local area connectivity, similar to node degree in graph theory. While node degree is the number of edges that a node is connected to, here, to account for the reliability between two interacting proteins as well as the differential gene expression status of the genes corresponding to these proteins, we defined local area connectivity for a given protein, n (c_n) as the product of the scaled differential expression status of the protein and the sum of the product of the differential expression status of its interacting partners. The strength of their interaction as shown in equation (4.3):

$$c_n = \left(\frac{d_n + 1.2}{4N}\right) x \sum_{i=1}^{N} e_i(d_i + 1.2) \tag{4.3}$$

where d_n and d_i represent the gene differential expression status of proteins n and I respectively while e_i is the reliability of the interaction between proteins n and i, and N is the total number of interacting partners for protein n.

The concept of guilty by association states that if two proteins interact together, then they are more likely to perform similar functions [143]. The concept, therefore, implies that if one of the proteins is known to be associated with a disease, then there is a high probability that another interacting partner is also likely to be associated with that disease. Using this concept, we modified equation (5-3) to include prior knowledge on whether a protein interacts with another protein that has been implicated in colorectal cancer or any other form of cancer using the COSMIC's cancer gene census [268]. Prior knowledge is defined as the ratio of the number of interacting proteins found in the cancer gene census list to the total number of interacting partners that a protein has, as shown in equation (4.4):

$$c_n' = \alpha \, c_n \tag{4.4}$$

where s is the number of interacting proteins that are part of the cancer gene census and α indicates whether n is listed on the cancer gene census list and $\alpha \in (0,1)$.

For each of the aforementioned networks generated, local area connectivity was calculated for each protein in the network to form a matrix LAC_m where the rows indicate the proteins and their local area connectivity values in each sample is represented as a column. Using SW480 as the reference dataset, we measured variability in the local area connectivity of each gene by computing the z-score as in equation (4.5):

$$z_n = \frac{x_n - \mu_{ns}}{\sigma_{ns}} \tag{4.5}$$

where z_n is the z-score for gene 'n' and x_n is the LAC score for gene 'n' in the SW480 cell line while μ_{ns} and σ_{ns} are the mean and standard deviation respectively of gene 'n' in either the TCGA patient samples or cell lines.

4.5.4. Enrichment analysis

Enrichment analysis was performed using FunRich [242], a functional enrichment analysis tool. A list of all proteins identified as having a change in local area connectivity in comparison to the reference dataset was imported into FunRich where functional and pathway enrichment analysis was performed.

4.5.5. Statistical analysis

SciPy’s [269] independent samples t-test was used to perform the statistical analysis and the differences were considered to be significant if the P value < 0.05.

Chapter 5

Integration of heterogeneous 'omics' data using semi-supervised network labelling to identify essential genes in colorectal cancer.

This chapter has been peer-reviewed and published in the Journal of *Computers and Electrical Engineering* (Chisanga, et al. [270]) and is presented here as a manuscript.

David Chisanga[a], Shivakumar Keerthikumar[b], Suresh Mathivanan[b] and Naveen Chilamkurti[a*]

[a]Department of Computer Science and Information Technology, La Trobe University, Bundoora, Victoria, 3086, Australia

[b]Department of Biochemistry and Genetics, La Trobe Institute for Molecular Science, La Trobe University, Melbourne, Victoria, 3086, Australia

*To whom correspondence should be addressed

Dr Naveen Chilamkurti

Department of Computer Science and Information Technology,

La Trobe University,

Bundoora, Victoria 3086, Australia

Tel: + 61 03 9479 1269

Fax: +61 03 9479 3060

Email: N.Chilamkurti@latrobe.edu.au

5.1. Abstract

Colorectal cancer (CRC) is the third most common form of cancer and has the fourth highest mortality rate in the world. To understand the origin and progression of this disease, biomedical researchers undertake global analyses of omics data of CRC patient samples and representative cell lines. However, due to the heterogeneity and high dimensionality nature of omics data, traditional tools for analysing this sort of data are inadequate, and the heterogeneous nature of cancer makes the process of identifying essential genes very difficult. This work uses network theory-based methods to address the problem of high dimensionality in omics datasets and applies network propagation to address the problem of heterogeneity in both omics datasets and cancer in identifying the essential genes in CRC. The method successfully identifies known essential genes in CRC as well as a new set of genes that are likely to be essential in the study of CRC.

Keywords: Proteomics, Genomics, Networks, Colorectal Cancer, Biomarkers, Machine Learning.

5.1. Introduction

Network theory, the study of how complex systems interact, is widely applied in fields such as computer networks, social networks, and interactome networks in systems biology [86]. Network metrics such as node degree are often used to prioritise nodes within a network. Similarly, one of the primary goals in cancer research is the identification of biomarkers or essential genes that can be used to understand the development or progression of a specific cancer type such as colorectal cancer (CRC).

To prioritise these genes, researchers often study the complex interactions between the numerous molecules within cells such as proteins, deoxyribonucleic acid (DNA), ribonucleic acid (RNA) and other small molecules which are obtained from the global profiling of patient samples, in addition to representative cell lines at multiple layers. These layers constitute what is today referred to as omics data [271] and consist of the transcriptome [272], genome [273], epigenome [274], proteome [275] and metabolome [276]. The interactions, on the other hand, are collectively known as interactome networks and provide a global picture of how molecular interactions influence cellular behaviour, an example being protein-protein interactions (PPI).

Omics data is highly dimensional in nature. Coupled with this is the heterogeneity of cancer whereby two individuals with the same type of cancer may have a different set of biomarkers. This makes identifying and prioritising cancer-related genes a challenging and daunting task that cannot be achieved using traditional statistical methods. As such, network theory provides a means by which complexity in such instances can be used to model the cellular system behaviour. Barabási et al. in [8] provide a summation of how network-based metrics can be applied in associating omics-related molecules to disease. Other work in [42, 43, 82, 90, 113, 138, 277-283] applied network-based methods in areas such as identifying and associating genes to disease and identifying drug targets in various cancer types. In [279, 284, 285], integrated network-based methods with machine learning techniques are used not only to reduce the dimensionality of omics data but to also build models that can use this data to predict genes associated with disease as well as classify multiple cancer types or tumour types. While the integration of omics data with networks has been gaining momentum over the years, a typical recurring theme in most of these researches has been the use of a single type of omics data as opposed to integrating the various types of omics data which are heterogeneous in nature.

In this chapter, building on the works discussed in Chapters 3 and 4 we used an integrated approach to identify essential genes in CRC, a type of cancer that originates in the bowel, is the third most common form of cancer and has the fourth highest cancer mortality rate in the world [13]. Unlike in Chapter 4 where we focused on the effect of APC mutations on the topology of the PPI network, here, we sought to understand the collective effect of all known gene mutations across a range of samples on the topology of the PPI network. Using the "central dogma of molecular biology" [286], we hypothesised that the mutation status and differential expression status of an individual gene has the potential effect on the expression of the protein that it codes for which in turn affects the global PPI network.

Our method employs a semi-supervised learning algorithm to propagate heterogeneous omics data into a PPI network and computes the likelihood distance of proteins from other proteins in the network whose corresponding genes are either mutated or differentially expressed. This was followed by a downstream enrichment analysis to validate and understand the role of the predicted potential essential genes in CRC.

5.2. Materials and methods

5.2.1. Proteomics data

We used proteomics and genomics data as the input to our method. Proteomics data consisted of PPIs. Weighted PPIs were downloaded from HIPPIE Version 2.0 [267], an online web-based database resource for weighted PPIs. The weights in the interactions show the confidence in the interaction between two proteins and are calculated by the authors based on the amount and reliability of evidence supporting an interaction. The PPI dataset was then filtered to leave out interactions with a confidence score of 0 after which 16,728 number of unique proteins and 276,183 number of interactions remain. These were then assembled into a network using NetworkX [287], a package in Python for network manipulation and analysis.

5.2.2. Genomics data

Genomics data comprised gene somatic mutations and gene differential expression status for CRC patients and representative cell lines. Previously, we collated genomics data related to CRC into a web-based resource called the Colorectal Cancer Atlas [11]. It is this data together with The Cancer Genome Atlas (TCGA) [28] patient data obtained from COSMIC [263] that we used as the genomics input data to our method. Using the corresponding genes for the proteins identified above, we obtained gene mutation details of 564 CRC patients from TCGA.

From the mutation dataset, we then filtered out all silent mutations and for each gene with a mutation in each sample, we represented its status using a binary number (1 if a mutation was present and 0 if not present), regardless of the number of mutations in a gene in each sample. The mutation data were then represented as a matrix, **M** (16,728x564) with rows representing genes and columns representing a gene's mutation station status in each sample. The same was repeated for gene differential expression status in TCGA patient data. This was then represented as a matrix, **D**(16,728x564) with rows representing genes and columns representing the differential expression status of genes in each sample. The gene differential expression status was denoted 1 for under-regulated or up-regulated genes and 0 for genes not differentially expressed.

5.3. Theory/Calculation

To identify essential genes, we use a method that integrates the different datasets discussed in section 5.2, materials and methods. Figure 5-1 provides a summary of the approach taken in this work.

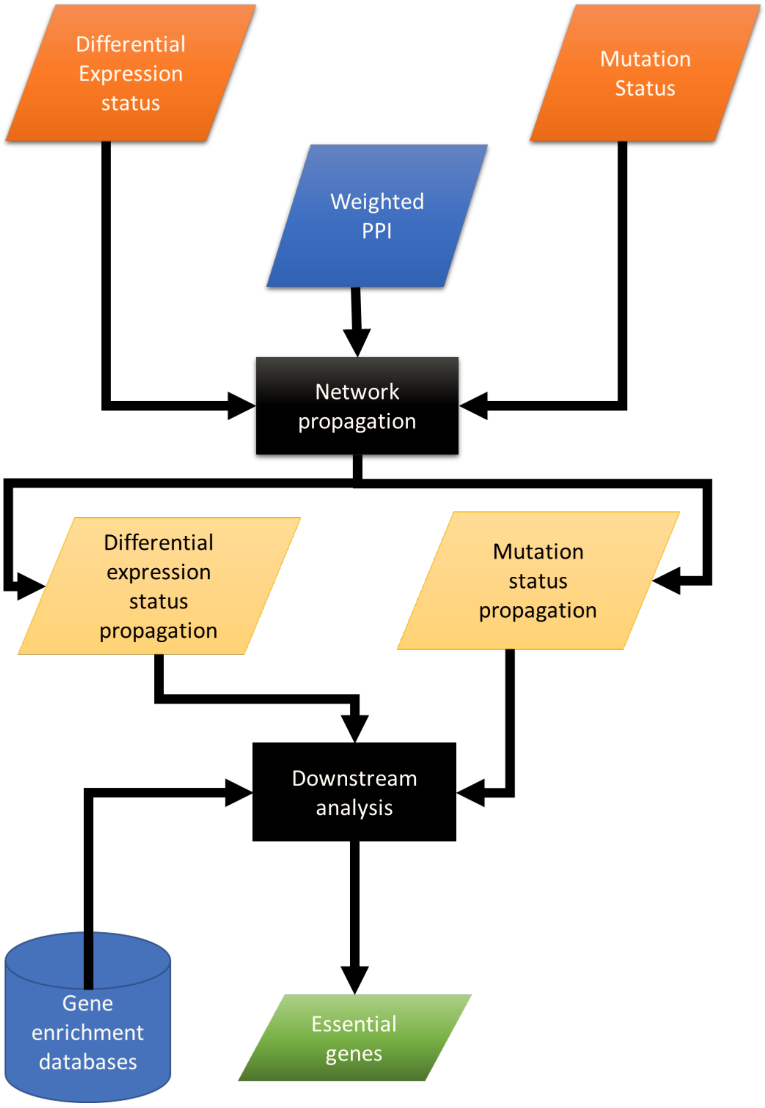

Differential Expression status
Mutation Status
Weighted PPI
Network propagation
Differential expression status propagation
Mutation status propagation
Downstream analysis
Gene enrichment databases
Essential genes

Figure 5-1: Architecture of model. Differential expression status and mutation propagation status were propagated through the network. The propagation results were then integrated together to form the features which were used in the further downstream analysis.

5.3.1. Disease gene prioritisation using network theory methods

A network or a graph is defined as a set of objects (nodes) linked together by lines (edges) [86]. A network is, therefore, represented as an ordered pair G=(V, E) where V is the set of nodes and E is the set of edges. By grouping a collection of objects as a set of nodes and using edges to represent relationships between these objects, researchers have used networks to reduce the complexity of large systems. Molecular networks in biology provide a global representation of the complex interactions between various molecules within a cell such as DNA, RNA and other small molecules.

When it comes to disease-gene prioritisation, many researchers use networks to associate genes with diseases. A naïve approach that is usually taken is to predict those genes that have neighbours associated with a disease as being more likely to be implicated in such a disease, that is, using the concept of "guilty by association". Such methods that implicate neighbours as having the likelihood of being associated with a disease include node degree as well as shortest path methods. However, these methods are prone to false positives because of the biases that exist in current molecular networks' datasets where proteins which are well studied tend to have more interactions than those that are not. Also, biological networks tend to obey the concept of the "small world" property where each node is reachable to another node through a series of links with other nodes and as such, the average number of hops needed to get to the furthest node from any given node is small [164].

5.3.2. Network propagation

Here, we used network propagation, a semi-supervised labelling algorithm first proposed by Zhou et al. [288] and further extended by Vanunu et al. [166] and Ruffalo et al. [289]. The objective was to determine the extent to which a gene's mutation status or differential expression status is propagated globally in a PPI network, and how it ultimately affects the topology of the network. The propagation results were then used to perform enrichment analysis to validate and determine roles played by the predicted essential genes in CRC. The input to the algorithm was a semi-labelled vector of gene mutation status M_v or differential expression status D_v, and a protein-protein interaction network G as shown in equation (5.1);

$$G(V, E, w) \tag{5.1}$$

where V is the set of proteins, E is the set of interactions and w is the set of interaction confidence scores (weight). The aim was to be able to determine the distance of the proteins in V (those that have not been labelled as either mutated or differentially expressed) from those that have been labelled as either mutated or differentially expressed.

For each node $v \varepsilon V$, we let N (v) be indicative of the direct neighbours of v in G. Let F: $V \rightarrow \Re$ be the propagation function where F(v) denotes the distance of a protein from those that are either differentially expressed or mutated. Let Y: $V \rightarrow [0,1]$ denote a prior knowledge function matching genes that are known to be differentially expressed or mutated as one (1) and zero (0) if not.

$$F(v) = \alpha\left[\sum_{\mu \in N(v)} F(u)w'(v,u)\right] + (1-\alpha)Y(v) \tag{5.2}$$

where w' is a [v]x[v] matrix and is a Laplacian normalised form of w as described below, the parameter $\alpha \in (0,1)$ weighs the relative importance of the two constraints discussed above, F and Y are vectors of size [n] where Y is the prior knowledge. Using the iterative procedure suggested by Zhou et al. [31], we use an iterative procedure to compute network propagation as in equation (5.3):

$$F^t = \alpha W' F^{t-1} + (1-\alpha)Y \tag{5.3}$$

where $F^1=Y$ and W' represents w'. The iterative algorithm can be described as a process where proteins for which prior genomic (mutated or differentially expressed) information exists iteratively pass on this information to their neighbouring nodes and every other node further propagates the information from the previous round to its neighbours repeatedly until convergence.

W' is a square matrix which represents the Laplacian normalisation of an [n]x[n] adjacency matrix W which is built from the set of confidence scores between interactions. We build an adjacency matrix W with a non-zero indicating an interaction between the two nodes and vice-versa. We then use Laplacian normalisation to get the matrix W' as shown in equation (5.4):

$$W' = D^{-1/2} W D^{-1/2} \tag{5.4}$$

where $D^{-1/2}$ is a diagonal matrix such that D (i, i) is the sum of row i of W.

After computation of the normalised weighted matrix W', for each sample in our data sets, we then iteratively computed the propagation scores for each of the nodes in the PPI. Vector Y was set as the prior knowledge vector where all the nodes whose corresponding genes, known to either be mutated or differentially expressed, were set to 1 and 0 otherwise. The propagation was computed separately by propagating node mutation status using the mutation status dataset as well as for the differential expression status dataset resulting in

P_m for mutation-based propagation scores and P_d for differentially expressed-based propagation scores. The propagation scores are then used to perform the following computations: propagation mean scores for genes in the samples, standard deviation, covariance which is then used to perform further downstream analysis to identify essential genes.

5.4. Results and discussion

5.4.1. Propagation of omics data

Network propagation of mutation status and that of differential expression status data is performed. Figure 5-2 shows the distribution of scores in TCGA samples. The figure also shows the relationships between the propagation scores against their corresponding status data. From this, it is shown that genes with a high-frequency rate of mutation or differential expression across samples are labelled with a propagation score close to their initial label in the prior knowledge dataset. This is further confirmed by the sensitivity of the algorithm, as shown in Table 5-1. The sensitivity is calculated by comparing the total number of correctly predicted/labelled genes against the total number of genes known a priori.

We hypothesise that genes with high mutation or differential propagation scores have a closer relationship to those genes that are either mutated or differentially expressed while those with low propagation scores are distant from the mutated or differentially expressed genes in the network. Based on the remaining filtered genes, we then pick the genes with propagation scores and perform enrichment analysis.

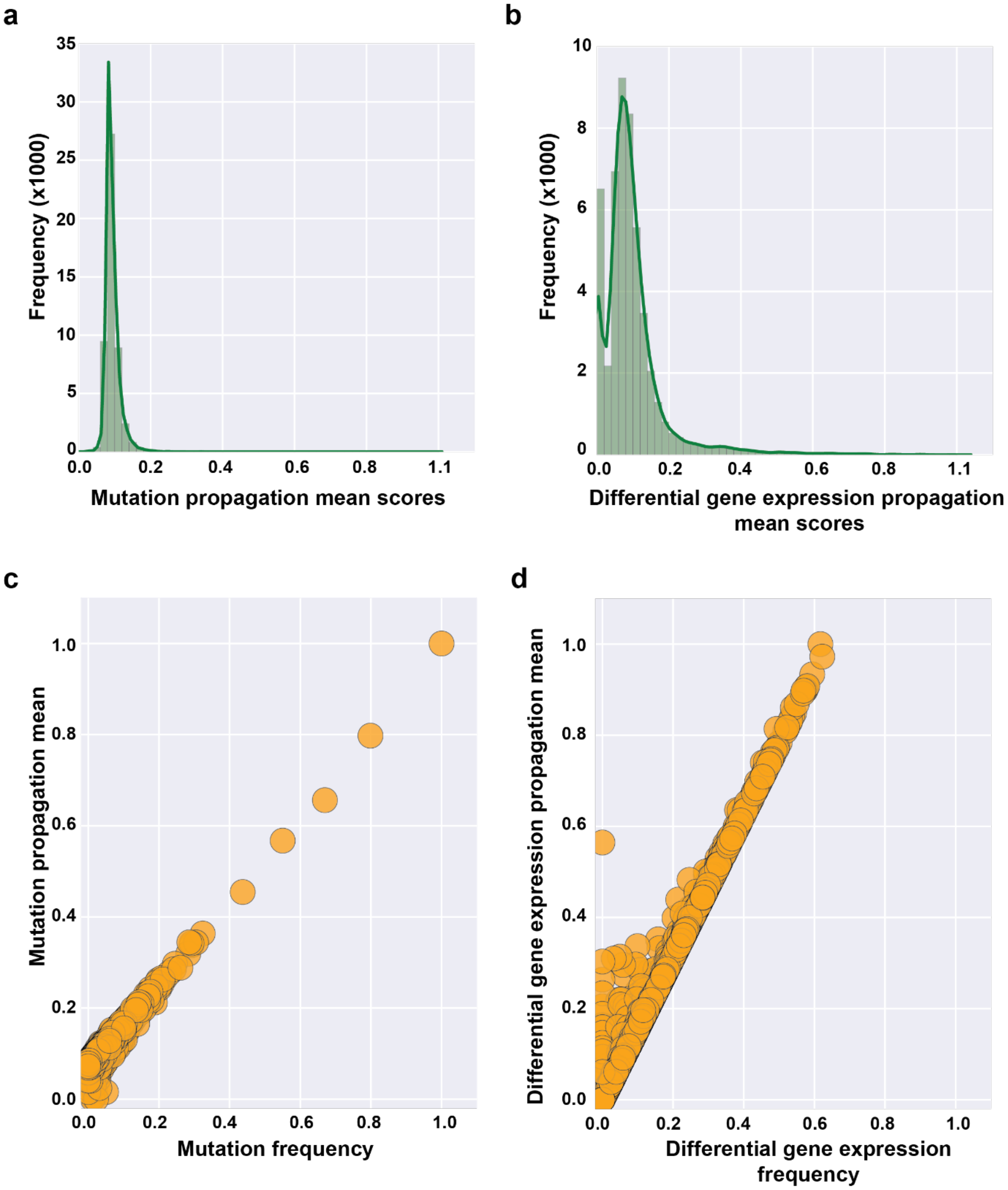


Figure 5-2: Summary of propagation scores in TCGA samples. (a) shows the distribution of mutation propagation scores (b) shows the differential expression status propagation scores (c) shows the relationship between the mean of mutation propagation scores against the mutation frequency (d) shows the relationship between the mean of differential expression status propagation scores against differential expression frequency

Table 5-1: Network propagation algorithm sensitivity scores. The sensitivity scores are used to measure consistency of network propagation in correctly labelling known genes as having high propagation scores similar to their previous labels

	Mutation	Differential expression
Number of correct labels	104505	565582
Number of incorrect labels	5	0
Sensitivity	0.999≈1	1.0

5.4.2. Enrichment analysis of mutation status propagation scores

To understand the relevance of the propagation results to CRC, we performed enrichment analysis on the propagation results using FunRich [242]. In Figure 5-3 (a) and (b), enrichment analysis of the genes with high mean mutation status propagation scores reveal that these genes are highly enriched in several cancers in the COSMIC database, furthermore, of these, it is found that 47 are also part of the COSMIC cancer gene census [268], as shown in Table 5-2.

Table 5-2: Genes found in COSMIC cancer gene census from propagation score

Genes from mutation propagation scores	Genes from differential expression propagation scores
AKAP9; ARID1A; ASXL1; ATM; ATP2B3; ATRX; BCL9L; BCORL1; BRAF; CASC5; CHD4; CIITA; FAT1; FAT4; FBXW7; GNAS; HLA-A; MT2A; KMT2D; KRAS; LIFR; LRP1B; MED12; MN1; MTOR; MYH11; NCOR2; NF1; NRAS; NRG1; PBRM1; PDE4DIP; PIK3CA; POLE; PREX2; PTPRT; RBM15; RNF213; RNF43; ROS1; RUNX1T1; SALL4; SMAD4; SPECC1; TCF7L2; TPR; ZFHX3	ASXL1; CUX1; ERCC5; MAP2K4; MYC; NONO; PHF6; PLCG1; RAD21; RB1; SMAD2; SMAD4; SRC; SS18; SS18L1; STAG2; TFE3; TOP1; UBR5; ZMYM

In addition, we also performed the biological process and molecular function enrichment to determine processes and functions most likely to be affected by the genes with high mutation status propagation scores as shown in Figure 5-3 (c) and (d) respectively. Of interest to us from the biological processes were homophilic cell adhesion and cell adhesion, as in [290] it is shown that these two processes play an important role in contact inhibition. Contact inhibition is cellular changes that lead to the termination of cell migration and proliferation because of signals transduced when one cell comes into physical contact with another cell. Nonetheless, in tumour microenvironments, it is shown that contact inhibition is lost due to the molecular changes in cell-cell adhesion, this, in turn, leads to cell proliferation and/or migration. This, therefore, means that changes in cell adhesion properties in cancer micro tumour environment play a key role in cancer progression and metastasis [16, 291]. Genes enriched in the two pathways are also shown in Table 5-3.

Table 5-3: Biological Process enrichment of genes with high mutation propagation scores

Biological process	**Enriched genes**
Homophilic cell adhesion	FAT3; ROBO2; FAT4; SDK1; ROBO1; DCHS2; PCDHA12; DSCAM; PCDHA7; PTPRT; PCDH10; FAT1; PCDHA6; TENM3; CELSR1; DCHS1; PCDH9; PCDH11X; CELSR2; CDH18; PCDHB3; PCDH20; PCDHB8; PCDHA3; SDK2; CDH23; PCDHA11; PCDH17; PCDHA2; PCDHA9; PCD-HGB2; PCDHA5; DSCAML1; PCDHGA11; PCDHA4; PCDHA10;
Homophilic cell adhesion	FAT3; ROBO2; FAT4; SDK1; ROBO1; DCHS2; PCDHA12; DSCAM; PCDHA7; PTPRT; PCDH10; FAT1; PCDHA6; TENM3; CELSR1; DCHS1; PCDH9; PCDH11X; CELSR2; CDH18; PCDHB3; PCDH20; PCDHB8; PCDHA3; SDK2; CDH23; PCDHA11; PCDH17; PCDHA2; PCDHA9; PCD-HGB2; PCDHA5; DSCAML1; PCDHGA11; PCDHA4; PCDHA10;

On the other hand, from the molecular function enrichment, it was found that genes that had high mutation propagation scores were also enriched in calcium ion binding and ATP binding molecular functions, as shown in Table 6-1. Calcium ion binding is part of the calcium cell signalling pathways whereby proteins bind to the Ca^{2+} ion. This pathway is important in regulating various cellular processes. A dysregulation of calcium ion binding function in cancer cells has been linked to the hyperpolarisation of tumour cells and impacts cancer cell proliferation and metastasis [292, 293]. In addition, related to calcium ion binding functionality is the ATP (adenosine triphosphate) binding function which acts as a source of energy needed by the ATP-binding cassette transporters to translocate substrates across membranes. The increased expression of ATP-binding cassette members has been shown to play a role in multi-drug resistance in diseases such as cancer [294-296]. These results, therefore, demonstrate that by propagating mutation status across the network, we can prioritise high scoring genes and their associated pathways and processes that are most likely to be affected by mutated counterparts.

Table 5-4: Molecular function enrichment of genes with high mutation propagation scores

Molecular function	**Enriched genes**
Calcium ion binding	PROC; TTN; FAT3; PCLO; DST; CACNA1B; NRXN1; FAT4; RYR2; FLG; DCHS2; PCDHA12; MEGF8; CACNA1E; FBN2; TENM2; CDHA7; LRP1B; BRAF; TCHH; ADGRL3; RYR1; PCDH10; GPR98; FAT1; PCDHA6; SLIT3; HMCN1; RYR3; CELSR1; SPTA1; CUBN; FBN3; DCHS1; PCDH9; PCDH11X; CELSR2; CDH18; FBN1; VCAN; PCDHB3; TBC1D9; DNAH7; HRNR; MEGF6; TPO; PCDH20; SLC25A12; PCDHB8; SLC25A23; PCDHA3; CDH23; PCDHA11; PKDREJ; PCDH17; PCDHA2; PCDHA9; LTBP3; PCDHGB2; LRP2; PCDHA5; STAB1; PCDHGA11; EFCAB6; ITPR1; ASTN2; LTBP4; PCDHA4; TNNC1; FSTL5; PLCH2; PCDHA10; MATN4;

ATP binding	TTN; PIK3CA; ABCA13; CACNA1B; OBSCN; DNAH10; DNAH14; DNAH2; KIF26B; ABCA7; CHD4; BRAF; ATP10A; RYR1; HELZ2; ATRX; DNAH5; DNAH9; MYH11; NLRP7; MDN1; DNAH8; EP400; LATS2; NAV3; TTBK1; MYH13; MYO18B; DNAH1; ACACB; ATM; DNAH11; ATP2B4; DNA2; SPEG; MYO3A; EPHB1; NWD1; SRCAP; DNAH7; ATP8B2; PHA3; ADCY8; WNK1; NLRP4; KIF1A; CIITA; CHD6; KIF4B; ATP13A3; ATP2B3; ROS1; NLRX1; SETX; ATP7A;
	SCN8A; LRRK2; DNAH6; ATP8B1; ABCA4; SMARCA2; DNAH3; ABCA12; MYO15A; NLRP5; MTOR; ATP11A; SMC1B; TTLL11; EPHA10; NRK; MYH3;

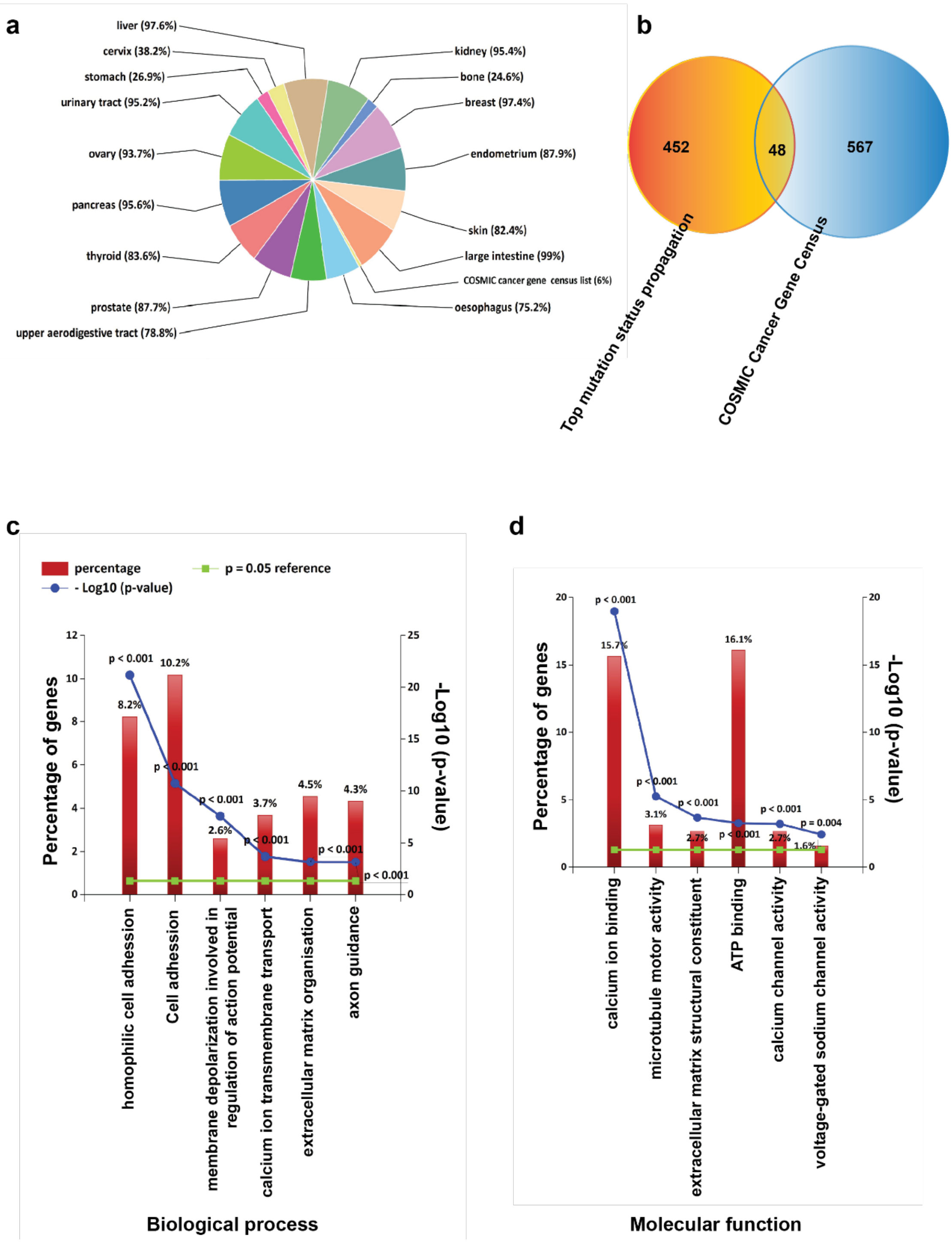


Figure 5-3: Enrichment analysis of genes with high mutation status propagation scores. (a) shows that genes with high mutation status propagation scores are highly enriched in different types of cancers in COSMIC (b) shows that 47 genes short-listed from the high mutation propagation scores are also found in the COSMIC census gene lists, (c) shows the biological process of the genes with high mutation status propagation scores, (d) shows the molecular function enrichment of genes high propagation scores

5.4.3. Enrichment analysis of differential expression status propagation scores

We also performed enrichment analysis for genes with high mean differential expression status propagation scores, as shown in Figure 5-4 and Tables 6.2, 6.5 and 6.6. The results show that similar to the mutation status propagation enrichment previously discussed, genes with high differential expression status propagation scores are highly enriched in various types of cancer from the COSMIC database. A comparison against COSMIC's cancer gene census shows that 20 of these genes are also found on the census list and the biological process and molecular function enrichments are not as significant as above. Nonetheless, of the significantly enriched molecular functions, dysregulation in Ubiquitin-specific protease activity has been shown to be associated with cancer [297-299], and members have been studied as potential drug targets for the treatment of cancer [300].

Table 5-5: Biological process enrichment of genes with high differential expression propagation scores

Biological process	Enriched genes
Regulation of nucleobase, nucleoside, nucleotide and nucleic acid metabolism	DDX27; NELFCD; TAF4; NCOA6; GMEB2; TCFL5; RPRD1B; PLAGL2; RBM39; RALY; DHX35; CSTF1; PCIF1; NCOA5; SUPT20H; HNF4A; RNF6; ASXL1; RNF113A; PHF20; ADNP; PDRG1; ZMYND8; MRGBP; TGIF2; GTF2F2; JADE3; NUFIP1; ZGPAT; ZFP64; FTSJ1; GZF1; PAN3; NONO; PQBP1; PARP4; UCKL1; SAP18; DKC1; UPF3A; NKRF; GTF3A; XRN2; PHF8; HNRNPH2; PABPC1; TRMT2B; ZNF696; HSF1; WBP4; ERCC5; ZNF623; CHRAC1; MYBL2; MAF1; ZNF34; PRICKLE3; ZHX3; RAD21; ZBTB33; TOP1MT; FAM50A; POLA1; UTP14C; TAF2; DCAF13; ZNF7; CRNKL1; TFDP1; DIS3; MED30; GTF2E2; RBM41; HUWE1; CNOT7; ZNF217; TOP1; UPF3B; TDRD3; MORF4L2; V39H1; CTPS2; GRHL2; HDAC6; PDS5B; HDAC8; PUF60; ZNF706; SCML2; ZFP41; INTS6; DSCC1; RBMX2; ZNF41; ZC3H13; SS18; DNMT3B; TFE3; POLR3D; HMGB1; PHF6; E2F1; POLR1D; KRBOX4; ASH2L; RB1; ZNF335; MBD1; CUX1; THOC2; ZNF337; CBFA2T2; SMAD4; MECP2; MYC; ID1; ZMYM2; ZSCAN25; ZMIZ2; ZC3H3; ZNF24; ZNF250;

Table 5-6: Molecular function enrichment of genes with high differential expression propagation scores

Molecular function	**Enriched genes**
Transcription regulator activity	NELFCD; NCOA6; PLAGL2; RBM39; PCIF1; NCOA5; SUPT20H; HNF4A; RNF6; ASXL1; PDRG1; ZMYND8; MRGBP; SS18L1; JADE3; PQBP1; SAP18; SCAND1; UXT; NKRF; ZNF696; MAF1; PRICKLE3; ZHX3; ZBTB33; MED30; SMAD2; CNOT7; MORF4L2; HDAC6; HDAC8; SCML2; ZC3H13; SS18; PHF6; RB1; ZNF335; MBD1; CUX1; ID1; ZMYM2; ZMIZ2; ZNF24; ZNF250;
Ubiquitin-specific pro-tease activity	CUL4A; RNF114; LNX2; ITCH; NEURL2; UBE2C; RNF219; TMEM189; COPS5; UBR5; UCHL3; UBL3; FBXL3; CUL1; UBL4A; SUGT1; UBE2D4; USP12; UBE2A; PSMD10; RNF216; PJA1; SCRIB; UBE2G1;

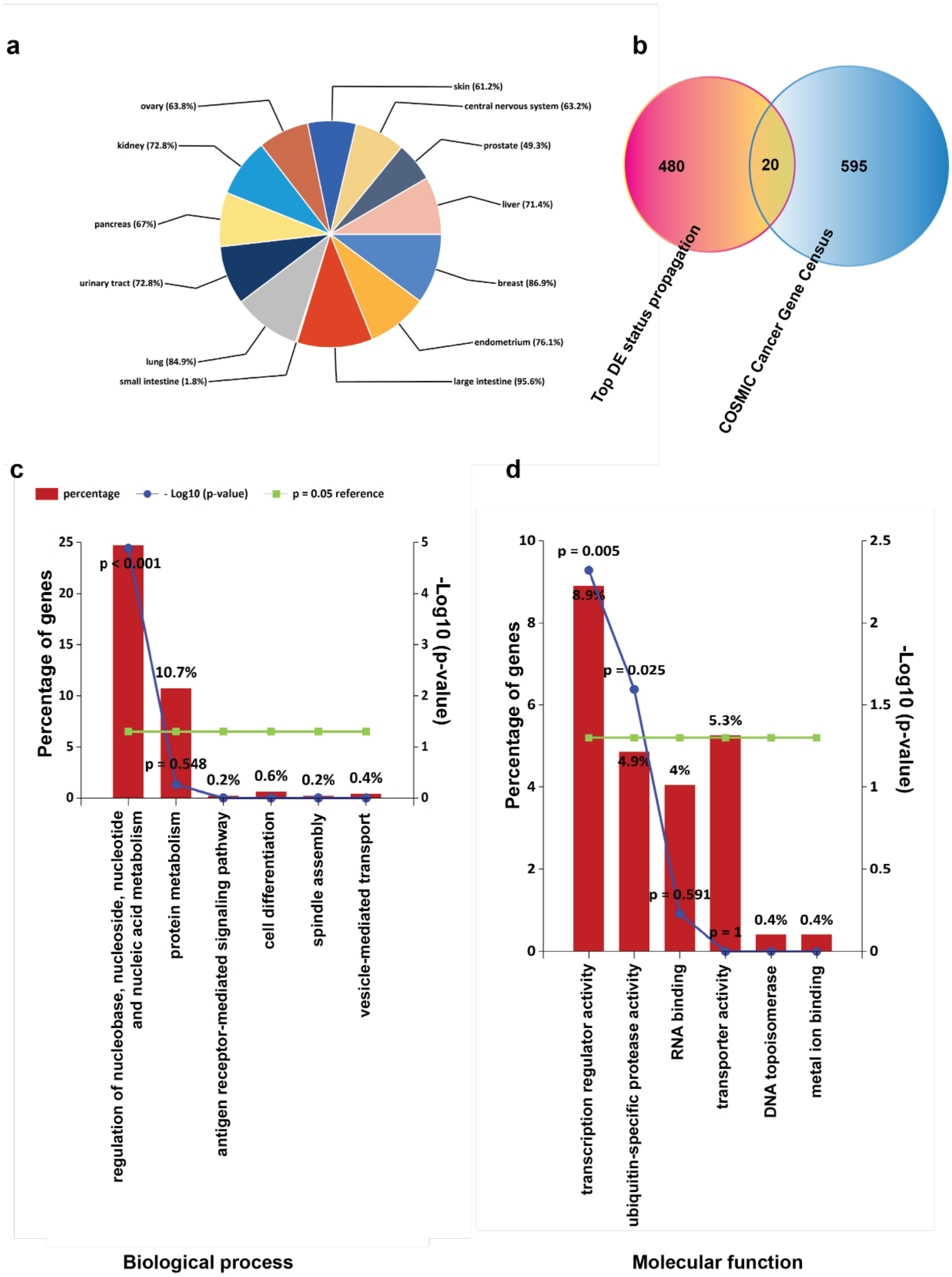


Figure 5-4: Enrichment analysis of genes with high differential expression status propagation scores. (a) shows that genes with high differential expression status propagation scores are highly enriched in various forms of cancers in COSMIC (b) shows that 20 genes short-listed from the high differential expression scores are also found in the COSMIC census gene lists, (c) shows that genes are only significantly enriched in one

biological process, (d) shows that genes with high differential scores are only enriched in two molecular functions

5.4.4. Linking mutation status and differential expression status scores

From the two lists of genes with high propagation scores, we filter for genes that appear in both lists, obtaining a set of 8 genes as shown in Figure 5-5, two of which are also enriched in COSMIC cancer gene census. These genes are considered as being close to both mutated and differentially expressed genes in the network. The following is the list of the identified genes: RALY, **ASXL1**, DIDO1, AP11A, ZC3H13, UGGT2, CCAR2 and **SMAD4**. ASXL1 and SMAD4 are known to be driver genes in cancer and are part of the COSMIC cancer gene census dataset. For instance, ASXL1 has been implicated in myelodysplastic syndrome (MDS) and chronic myelomonocytic leukaemia (CML) while SMAD4 has been implicated in the following cancer types: colorectal, pancreatic, and small intestine. On the other hand, a literature search of the remaining six genes shows that they have also been implicated in some of form of cancer with varying roles ranging from resistance, metastasis and cell proliferation. For example, RALY is a gene that codes for the protein RNA-binding protein and in [301] has been implicated to play a role in the development of drug resistance in CRC; DIDO1 is a gene which codes for the protein death inducer-obliterator and is involved in apoptosis or cell death and has been found to affect cell viability and anchorage in CRC cells [302]; and CCAR2 has been implicated in other forms of cancer [303].

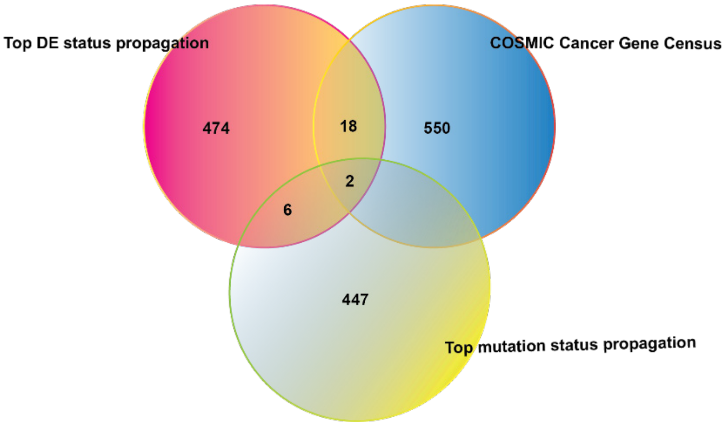


Figure 5-5: The Venn diagram shows the genes found to be closer to genes that are differentially expressed and have a mutation. The Venn diagram also shows the genes that were found on the COSMIC's cancer gene census.

5.5. Conclusion

The rate at which omics data is generated has over the years been rising substantially and is expected to rise further due to the continued decline in the cost and the advancements in high-throughput technologies such as next-generation sequencing technologies. As such traditional statistical methods can no longer be relied upon as a way of analysing such gigantic amounts of data. Network analysis, the evaluation of how nodes relate to one another coupled with new machine learning methods, has over the years become an integral tool for analysing high throughput data such as omics data.

In this chapter, we demonstrated how heterogeneous omics datasets can be integrated by use of network-based methods and how features can be prioritised using a semi-supervised technique coupled with further downstream analysis. We found that the method successfully identified the essential genes in CRC. Further, we also identified new genes that may play a role CRC in the development and progression of cancer. However, the genes that were predicted in this paper need further experimental validation to understand their specific roles in CRC. In addition, this study was limited by the lack of vast amounts of paired wild-type and mutant data, this, in turn, made it difficult to further explore our findings and incorporate soft computing techniques. Future works include fine-tuning the current model and validating the predicted genes using wet laboratory experiments. We also plan on incorporating new machine learning techniques such as deep learning using neural networks.

Chapter 6

Physical coherence and network analysis to identify novel regulators of exosome biogenesis.

This chapter is in preparation as a manuscript for submission to a peer-reviewed journal.

6.1. Abstract

Exosomes are small membranous vesicles of endocytic origin with a diameter of 30-150nm. Exosomes have been implicated in a range of biological functions such as intercellular communication through the transmission of macromolecules such as proteins, nucleic acids and lipids, as well as in the pathogenesis and progression of diseases such as cancer. Therefore, there has been growing interest in understanding the biogenesis, functionality, and applications of exosomes in both physiological and pathological conditions.

The biogenesis of exosomes has long been associated with the endosomal sorting complex required for transport (ESCRT) machinery, together with other accessory proteins. However, the mechanisms behind exosome biogenesis are still poorly understood, and the proteins involved in the process of exosome biogenesis have not all been characterised. Here, we, therefore, attempt to identify novel proteins that regulate the process of exosome biogenesis through the ESCRT pathway and improve our understanding of exosome biogenesis. To achieve this, network analysis methods are applied to a protein-protein interaction (PPI) network of the ESCRT machinery. To counter the bias that exists in PPIs due to false positives stemming from experimental errors in the techniques used to identify them, we extend the network analysis method by using physical coherence, a technique that quantifies the connectedness of a PPI network due to topological changes. Using this technique, STAMBP and NEDD4 are predicted as potential novel regulators of exosome biogenesis. It was found that STAMBP increased the physical coherence of the ESCRT machinery network while NEDD4 reduced the physical coherence of the ESCRT machinery network. To validate our findings, SDCBP, a protein that has been previously shown to regulate exosome biogenesis was also found to change the physical coherence of the ESCRT machinery. Further analysis using CRISPR-Cas9-based knockout cells of NEDD4 and STAMBP confirms their active role in exosome biogenesis.

6.2. Background

6.2.1. Exosome biogenesis

Exosomes are small membranous vesicles of endocytic origin with a diameter in the range of 30-150 nm that are secreted under normal and pathological conditions [183, 194]. Exosomes have been implicated in a range of functions such as acting as a channel of communication between cells through the transmission of macromolecules such as proteins, nucleic acids and lipids. They have also been implicated in the development and progression of diseases such as cancer [198, 304]. In addition to their role of cellular communication, exosomes have also been implicated as potential vectors that can be used to carry and deliver drugs for therapeutic applications [305]. Because of these functions attributable to exosomes, there has been growing interest in the study of the biogenesis, functions and applications of exosomes in both physiological and pathological conditions.

The biogenesis of exosomes starts with the inward budding of endosomal membranes, resulting in the formation of intraluminal vesicles (ILVs) within the multivesicular bodies (MVBs). Upon maturation, the MVBs fuse with the plasma membrane and their contents are then secreted into the extracellular space as exosomes. This process is summarised in Figure 6-1. The mechanism by which exosome biogenesis takes place is however not yet fully understood. Currently, the ESCRT machinery together with other accessory proteins are thought to regulate exosome biogenesis [183, 205, 306]. The ESCRT machinery is made up of approximately 20 proteins which interact together to form four components: ESCRT-0, I, II and III. These components are linked together to form a network complex and work in a sequential order in association with other accessory proteins such as ALIX and VPS4 [204]. The ESCRT machinery components ESCRT-0 recognise ubiquitylated proteins in the endosomal membrane while ESCRT-I and II complexes are responsible for the sorting of cargo and inward budding of membranes, and the ESCRT-III component is responsible for vesicle scission [206].

In addition to the ESCRT machinery, other ESCRT-independent pathways have been shown to regulate exosome biogenesis. Examples of such pathways include those that are involved in tetraspanins, Rab GTPases and lipids. Nonetheless, it has been established that the ESCRT machinery is conserved in several organisms [307] and is therefore thought to regulate exosome biogenesis in various organisms. While there is ongoing research to understand the mechanisms behind exosome biogenesis and how the ESCRT machinery performs its role, not all the proteins that are involved in the process of regulating exosome biogenesis have been identified. Our aim here, therefore, is to identify proteins that can regulate the biogenesis of exosomes through the ESCRT pathway.

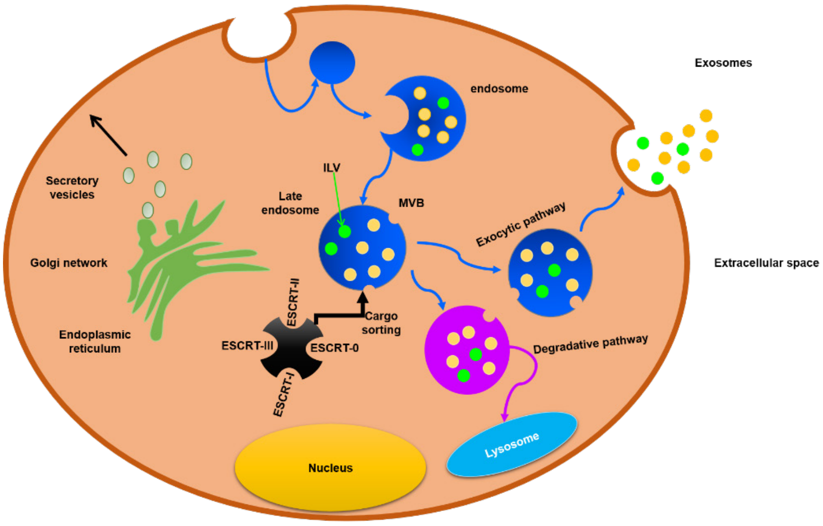


Figure 6-1: Exosome biogenesis and secretion. Exosomes are formed by the inward budding of the endosomal membranes. The ESCRT and other associated proteins such as ALIX and TSG101 are implicated in the sorting of cargo and the formation of the intraluminal vesicles (ILV). The MVBs then either merge with the plasma membrane and release their content into the extracellular space as exosomes or fuse with lysosomes [183].

6.2.2. Physical coherence and network analysis

To understand the interplay of the ESCRT pathway and its interacting partners in regulating exosome biogenesis, we apply network theory in this chapter. Network theory, the study of how objects interact with each other, has long been used in fields such as computer science and engineering [308], sociology [309], and physics [170] for visualisation and the reduction of complexity in various systems. In systems biology and medicine, network analysis methods are applied in areas such as drug target identification [310, 311], the prediction of protein function [312], and protein complex detection [313]. Other areas include the prediction of novel interactions and functional associations [314], the identification of disease sub-networks [315], disease biomarker identification [82], and the mapping of disease pathways. Network theory metrics such as degree centrality (DC), closeness centrality (CC) and betweenness centrality (BC) are commonly used in complex networks to identify essential objects within a network.

Here, network analysis methods are applied to a PPI network of the ESCRT machinery to predict novel proteins that are likely to regulate exosome biogenesis via the ESCRT pathway. In chapter 2, the role of PPIs in physiological conditions, together with their definitions, is discussed. PPIs provide a simplified global picture of the underlying complex functional make-up of the cell. Chapter 2 further discusses numerous studies that have been conducted to map PPIs and how they have been collated into several online databases such as the Human Protein Reference Database (HPRD). However, these mappings are prone to high false positives which emanate from errors in experimental techniques used to identify them, as well as technical and study biases [316]. Hence, frequently studied proteins such as those associated with diseases like cancer tend to have higher degree centralities compared to those that are less studied. Therefore, this implies that if using common network metrics, such as node degree, and proteins that are frequently and well-studied will often rank higher than those genes or proteins that are less studied.

It is against this background that, in order to identify any novel proteins of importance to exosome biogenesis, we need a method that eliminates the bias found in current PPIs. Sama and Huynen [9] and Oortveld, et al. [317] developed a technique called physical interaction enrichment (PIE), a method that quantifies the change in the physical cohesiveness of a PPI network given any topological change in a network. We apply this method in this study to measure the physical cohesiveness of an ESCRT PPI network when an ESCRT neighbouring protein is added to the network. The physical cohesiveness is computed by calculating the ratio of the number of interactions among a set of proteins when compared to the number of interactions in randomly generated networks. PIE eliminates the bias that exists in PPIs by normalising the degree centrality of proteins against randomly generated networks with similar degree distribution from an overall network representation.

6.3. Materials and Methods

Figure 4.1 provides a summary of the workflow that describes the steps and methods that were applied in predicting proteins that regulate exosome biogenesis through the ESCRT pathway.

6.3.1. Literature mining for ESCRT machinery proteins

Human proteins that make up the ESCRT machinery were collated and curated into their respective components. The human ESCRT proteins were mapped to their corresponding orthologs in three other model organisms (Worm (*Caenorhabditis elegans*), Fly (*Drosophila melanogaster*) and Yeast (*Saccharomyces cerevisiae*)), using an online database resource called InParanoid [318].

6.3.2. Construction of PPI Network

A PPI network is represented as an undirected graph G (V, E), where V is a set of nodes which denote proteins while E is a set of edges denoting the interactions between proteins.

Background PPI network: We downloaded interaction data from BioGRID version 3.4.134 [319] and HPRD release 9 [320]. From the BioGRID dataset, we obtained all interactions belonging to humans and all proteins with their corresponding orthologs in our chosen model organisms. These were combined with other unique interactions obtained from the HPRD dataset to form a comprehensive PPI network, referred to as the background database (G_{bg}).

Organism-specific ESCRT machinery network: Using the background PPI established above, we generated a human ESCRT PPI network (G_h). PPIs for the three other organisms were generated as follows: G_w for the worm, G_y for yeast and G_f for the fly. The four individual organisms' networks were then combined together ($G_h + G_f + G_y + G_w$) to form a bigger ESCRT network called the master network (G_m). By combining the four networks, we aimed at identifying as many potential interactions as possible between the ESCRT proteins across the four organisms and in turn maximise the number of ESCRT proteins' neighbours.

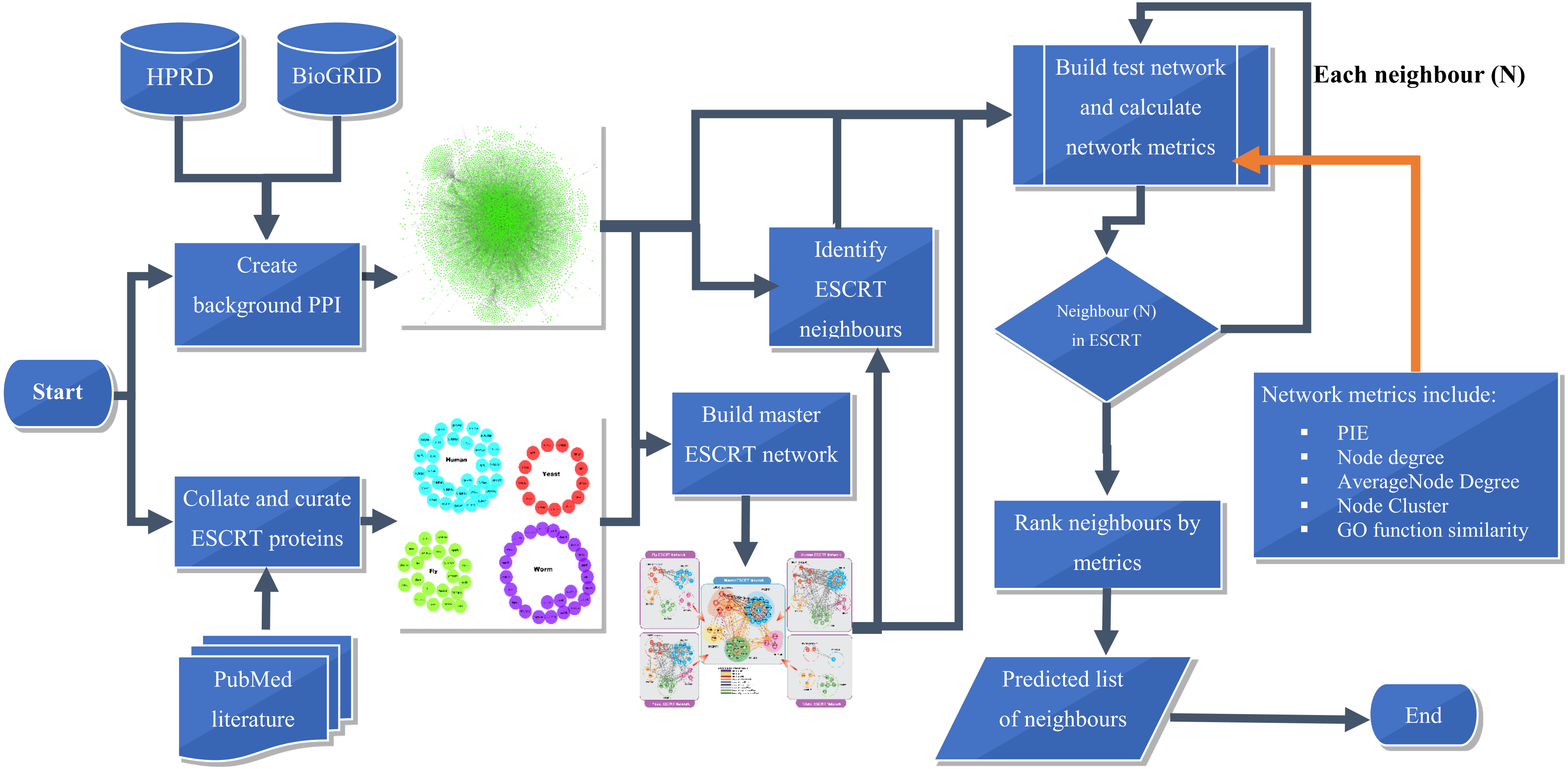


Figure 6-2: ESCRT neighbour prediction flowchart. PPI data was retrieved from HPRD and BioGRID and used to create a background PPI. ESCRT proteins were mined from the literature and were mapped to their corresponding orthologs in other organisms. The orthologs ESCRT networks and the ESCRT network in humans were combined to form the Master ESCRT network which was then used to mine for neighbouring proteins. Using the

neighbouring proteins and the ESCRT master network, test networks were generated for each neighbouring protein and used to calculate PIE and other network metrics

6.3.3. Physical coherence and network analysis

To determine which ESCRT machinery neighbouring proteins are likely to play a role in exosome biogenesis, we used physical interaction enrichment (PIE), a method proposed by Sama and Huynen [9]. To find essential proteins, network metrics such as node degree, betweenness centrality and page rank are generally used [8], however, such methods as discussed above are prone to false discoveries due to the bias that exists in PPI datasets. The physical cohesiveness of a network is calculated by normalising the median node degree of a network against the median node degree of a set of random networks that have the same node degree distribution as the network whose physical cohesiveness is being determined. In this study, to calculate PIE, a new test network is first created by combining the master ESCRT machinery (G_m) with all the interactions that neighbour n has with ESCRT proteins. The following gives an overview of the algorithm applied;

Algorithm

Step 1: For each ESCRT neighbouring protein n, a test network (G_n^t) was constructed by combining the master ESCRT PPI network (G_m) and the neighbouring protein's interactions where G_n^t is a subset of G_{bg} as shown in equation (6-1);

$$G_n^t \subset G_{bg} \tag{6-1}$$

Step 2: The node degree for each protein in the test network was calculated as the total number of other proteins that it interacts with, as in equation (6-2);

$$\text{degree} = \sum_{i=1}^{m} e \tag{6-2}$$

where e is the interaction between a given protein and its interacting partners m.

Step 3: Using the node degrees calculated from step 2, we established the node degree distribution for G_n^t where degree distribution was the number of nodes (proteins in this case) which have the same node degree from the network.

Step 4: Using the node degree distribution established in step 3, a set of all proteins (referred to as the background node-set) from G_{bg} that had the same degree distribution as G_n^t was obtained. Using this set of proteins, a set of random networks were generated such that for each random network generated, the number of proteins obtained from each node degree distribution was equal to the number of proteins in the corresponding node degree distribution obtained from step 3. For each test network G_n^t , a corresponding 1,000 random networks were generated based on a random combination of nodes from the background node set.

Step 5: The randomly generated networks were compared against the background node set using Wilcoxon test to ensure that the background node set was large enough to generate the random networks.

Step 6: Once the random networks were tested as outlined in step 5, the average node degree for each of the 1000 random networks was computed and arranged in ascending order (as set N_t) from where the median node degree was obtained and then used to calculate the PIE score for each neighbouring protein. PIE was calculated as shown in equation (6-3);

$$\text{PIE} = \frac{\mu_{degree}^t + 1}{median(N_t) + 1} \qquad (6\text{-}3)$$

where μ_{degree}^t is the average node degree for G_n^t and N_t is the set of average node degrees for the random networks.

Step 7: Each PIE score calculated in step 6 was then tested for significance as the ratio of the number of times the average node degree for G_n^t was greater than the average node degree for a randomly generated network in step 3 to the total number of random networks generated, as shown in equation (6-4);

$$\text{PIE} = \frac{n}{N} \qquad (6\text{-}4)$$

where n is the number of times the average node degree for G_n^t was greater than the average node degree for random networks and N is the total number of randomly generated networks.

This process was repeated for all the ESCRT machinery neighbouring proteins identified above after which the proteins were ranked based on the PIE score.

6.3.4. GO semantic function similarity

In addition to PIE, we also calculated the gene ontology (GO) semantic function similarity of the ESCRT neighbours to the ESCRT machinery. GO is a collaborative public database that offers a controlled vocabulary of terms that are used to describe gene products' functionality and the relationships between them. GO terms are classified into three main classes: molecular function, cellular component, and biological process. The GO framework is typically represented as a directed acyclic graph whereby each term has defined relationships to one or more other terms in the same domain or other domains [321], an example of which is shown in Figure 6-3.

Using GO terms and the three classes, we can deduce the functionalities of a given gene and its products. Likewise, here, in addition to physical coherence, we used GO function semantic similarity to determine how functionally related the ESCRT neighbouring proteins are to the ESCRT proteins.

GO semantic similarity is a method which measures the functional similarity between two genes or gene products based on the number of GO terms that co-occur between them. This method was first proposed by Wang, et al. [322]. GOSemSim, a package in R by Yu, et al. [323] was used to calculate the GO semantic similarity of each of the ESCRT neighbouring proteins n against each ESCRT protein and the results were added to set GO_n . Overall, GO semantic similarity for each neighbouring protein n was taken as the mean of all the scores in GO_n.

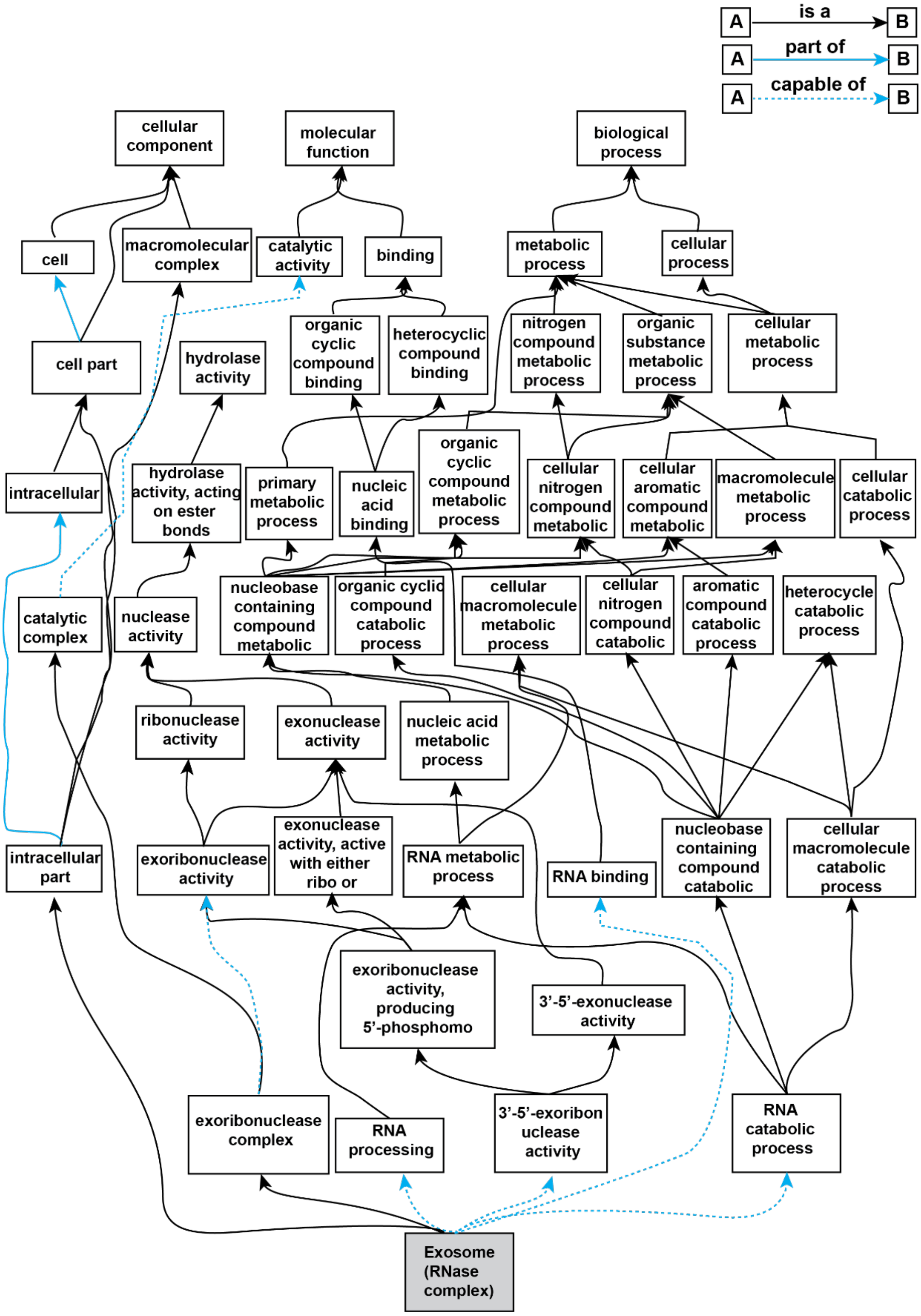

A
is a
B
A
part of
B
A
capable of
B
cellular component
molecular function
biological process
cell
macromolecular complex
catalytic activity
binding
metabolic process
cellular process
cell part
hydrolase activity
organic cyclic compound binding
heterocyclic compound binding
nitrogen compound metabolic process
organic substance metabolic process
cellular metabolic process
intracellular
hydrolase activity, acting on ester bonds
primary metabolic process
nucleic acid binding
organic cyclic compound metabolic process
cellular nitrogen compound metabolic
cellular aromatic compound metabolic
macromolecule metabolic process
cellular catabolic process
catalytic complex
nuclease activity
nucleobase containing compound metabolic
organic cyclic compound catabolic process
cellular macromolecule metabolic process
cellular nitrogen compound catabolic
aromatic compound catabolic process
heterocycle catabolic process
ribonuclease activity
exonuclease activity
nucleic acid metabolic process
intracellular part
exoribonuclease activity
exonuclease activity, active with either ribo or
RNA metabolic process
RNA binding
nucleobase containing compound catabolic
cellular macromolecule catabolic process
exoribonuclease activity, producing 5’-phosphomo
3’-5’-exonuclease activity
exoribonuclease complex
RNA processing
3’-5’-exoribon uclease activity
RNA catabolic process
Exosome (RNase complex)

Figure 6-3: Example of a GO tree structure. The start term is mapped to the other terms to which the term is related in a tree-like structure forming a parent-child relationship. Exosome here is the RNase complex and is not to be confused with extracellular exosomes secreted by cells.

6.3.5. Other network metrics

Other network metrics were used either to validate the significance of PIE scores or as a part of the PIE calculation methodology. Examples include the node degree, which is a measure of the number of interacting partners that a node has, average node degree which gives the average number of interactions a node has in each network, and node cluster coefficient which gives the fractions of possible triangles that go through that node.

6.4. Results and Discussion

6.4.1. Identification of ESCRT proteins and ESCRT orthologs

Using the process described in section 6.3.1 of the materials and methods, we identified 32 ESCRT proteins together with their corresponding orthologs in worm, fly and yeast, as shown in the table. Individual ESCRT PPIs for each of the four organisms were then constructed using the interaction as described above. The four networks are summarised in Figure 6-4, and it is shown that some of the interactions are conserved across the four organisms while others are specific to only some organisms.

Table 6-1: ESCRT and accessory proteins in Homo sapiens are mapped to their corresponding orthologs in *C. elegans*, *D. melanogaster* and *S. cerevisiae*

H. sapiens		**C. elegans**		**D. melanogaster**		**S. cerevisiae**	
Entrez Gene ID	Official Symbol	Entrez Gene ID	Official Symbol	Entrez Gene ID	Official Symbol	Entrez Gene ID	Official Symbol
27243	CHMP2A	174908	vps-2	43164	Vps2	853868	DID4
81553	FAM49A	174234	R07G3.8	39206	CG32066	-	-
51510	CHMP5	179242	vps-60	39964	CG6259	852097	VPS60
79643	CHMP6	171654	vps-20	37581	Vps20	855101	VPS20
57132	CHMP1B	171801	did-2	40036	Chmp1	853906	DID2
8027	STAM	172264	stam-1	34505	Stam	856387	HSE1

H. sapiens		**C. elegans**		**D. melanogaster**		**S. cerevisiae**	
Entrez Gene ID	Official Symbol	Entrez Gene ID	Official Symbol	Entrez Gene ID	Official Symbol	Entrez Gene ID	Official Symbol
128866	CHMP4B	174091	vps-32.1	35933	shrb	850712	SNF7
128866	CHMP4B	183288	vps-32.2	35933	shrb	850712	SNF7
51652	CHMP3	173863	vps-24	40542	Vps24	853825	VPS24
51652	CHMP3	3565940	rnh-1.0	40542	Vps24	853825	VPS24
51028	VPS36	179520	vps-36	39523	Vps36	851135	VPS36
51571	FAM49B	174234	R07G3.8	39206	CG32066	-	-
10015	PDCD6IP	176410	alx-1	43330	ALiX	854449	RIM20
5119	CHMP1A	-	-	-	-	853906	DID2
10254	STAM2	172264	stam-1	34505	Stam	856387	HSE1
55048	VPS37C	178944	vps-37	40624	Vps37B	850810	SRN2
11267	SNF8	175672	vps-22	42572	lsn	856105	SNF8
79720	VPS37B	178944	vps-37	40624	Vps37B	850810	SRN2
9525	VPS4B	189590	vps-4	32777	Vps4	856303	VPS4
7251	TSG101	182474	tsg-101	39881	TSG101	850349	STP22
27183	VPS4A	189590	vps-4	32777	Vps4	856303	VPS4
51534	VTA1	172528	T23G11.7	38204	CG7967	850878	VTA1
155382	VPS37D	-	-	-	-	850810	SRN2
9146	HGS	177617	hgrs-1	33458	Hrs	855739	VPS27
51160	VPS28	173229	vps-28	47408	Vps28	856040	VPS28
137492	VPS37A			31006	mod(r)	850810	SRN2
93343	MVB12A					853120	MVB12
84313	VPS25	173143	vps-25	35847	Vps25	853566	VPS25

H. sapiens		**C. elegans**		**D. melanogaster**		**S. cerevisiae**	
Entrez Gene ID	Official Symbol	Entrez Gene ID	Official Symbol	Entrez Gene ID	Official Symbol	Entrez Gene ID	Official Symbol
9798	IST1			38750	CG10103	855456	IST1
25978	CHMP2B	182050	C01A2.4	38599	CHMP2B	853868	DID4
29082	CHMP4A	183288	vps-32.2	35933	shrb	850712	SNF7
92421	CHMP4C	183288	vps-32.2	35933	shrb	850712	SNF7
91782	CHMP7		T24B8.2	174442	CG5498	853906	DID2

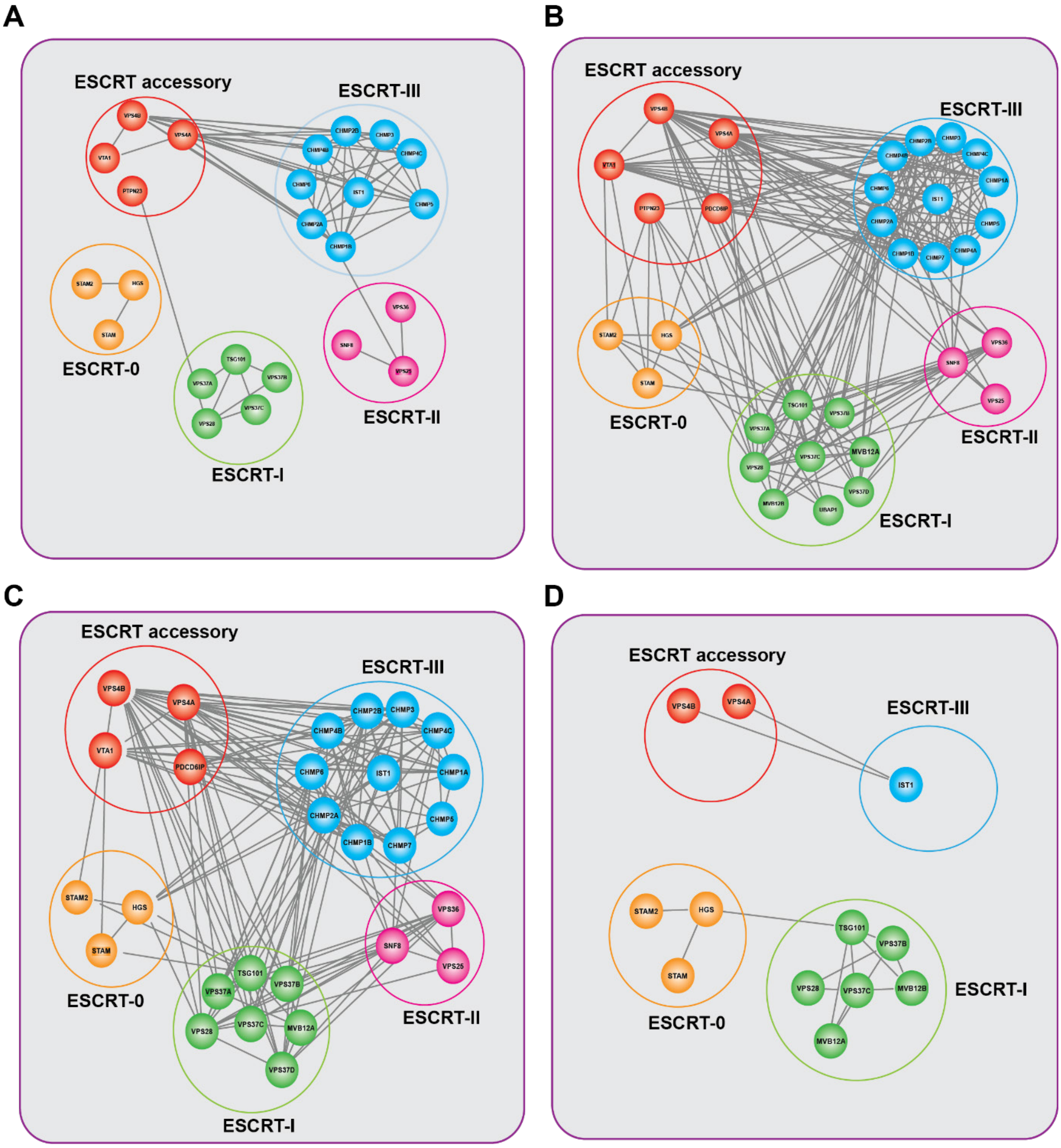


Figure 6-4: Organism-specific ESCRT networks. (A) ESCRT machinery network for Fly generated by mapping human ESCRT proteins to corresponding orthologs in Fly, (B) Human ESCRT machinery network generated from the background PPI, (C) Yeast ESCRT, similarly generated by mapping human ESCRT proteins to their corresponding orthologs in Yeast, (D) Worm ESCRT machinery network. By generating organism-specific ESCRT machinery networks, we aimed at maximising the number of neighbouring proteins that we could identify and check how conserved ESCRT protein interactions are across the four organisms.

6.4.2. Identification of ESCRT machinery neighbours

Individual model organisms' protein-protein interactions were generated based on the ESCRT interaction information from the background PPI database G_{bg}, and then combined to form the master network G_m as shown in Figure 6-5 (A). Using the master ESCRT machinery PPI, more than 1,800 neighbours were identified by searching in the background PPI database G_{bg} for those proteins that interact with any of the ESCRT proteins but are not part of the ESCRT machinery. For each neighbouring protein identified, its interactions with the ESCRT machinery network were mapped and a test network G_n^t generated, as shown by the example in Figure 6-5 (B) where STAMBP was mapped to G_m. The physical coherence score for each neighbouring protein test network was calculated and the proteins ranked by score.

6.4.3. Physical coherence changes in master network

To understand the contribution of each ESCRT protein to the physical coherence of the master network, we first computed the overall PIE score of the master ESCRT machinery G_m in relation to the background network G_{bg}, and the returned score was approximately 14.4 with a p-value=0.15. This value is taken as a reference point in the further analysis of physical coherence changes in the master network.

Furthermore, the contribution to the overall physical coherence of G_m by the ESCRT proteins was analysed by computing the PIE score of G_m when a given ESCRT protein and its interactions are removed from the master network. From the results, most of the ESCRT members did not show a significant change in the physical coherence of the network upon being removed from the master network. In contrast, HGS and STAM2 showed significant changes in the physical coherence of the master network upon being removed from the ESCRT network. The results are summarised in Figure 6-6 (A) which compares the physical coherence between PIE scores for the ESCRT machinery with and without a given ESCRT protein. From the results, we therefore hypothesised that the ESCRT machinery is highly coherent and that any protein that significantly changes the physical coherence of the network has a high probability of regulating exosome biogenesis.

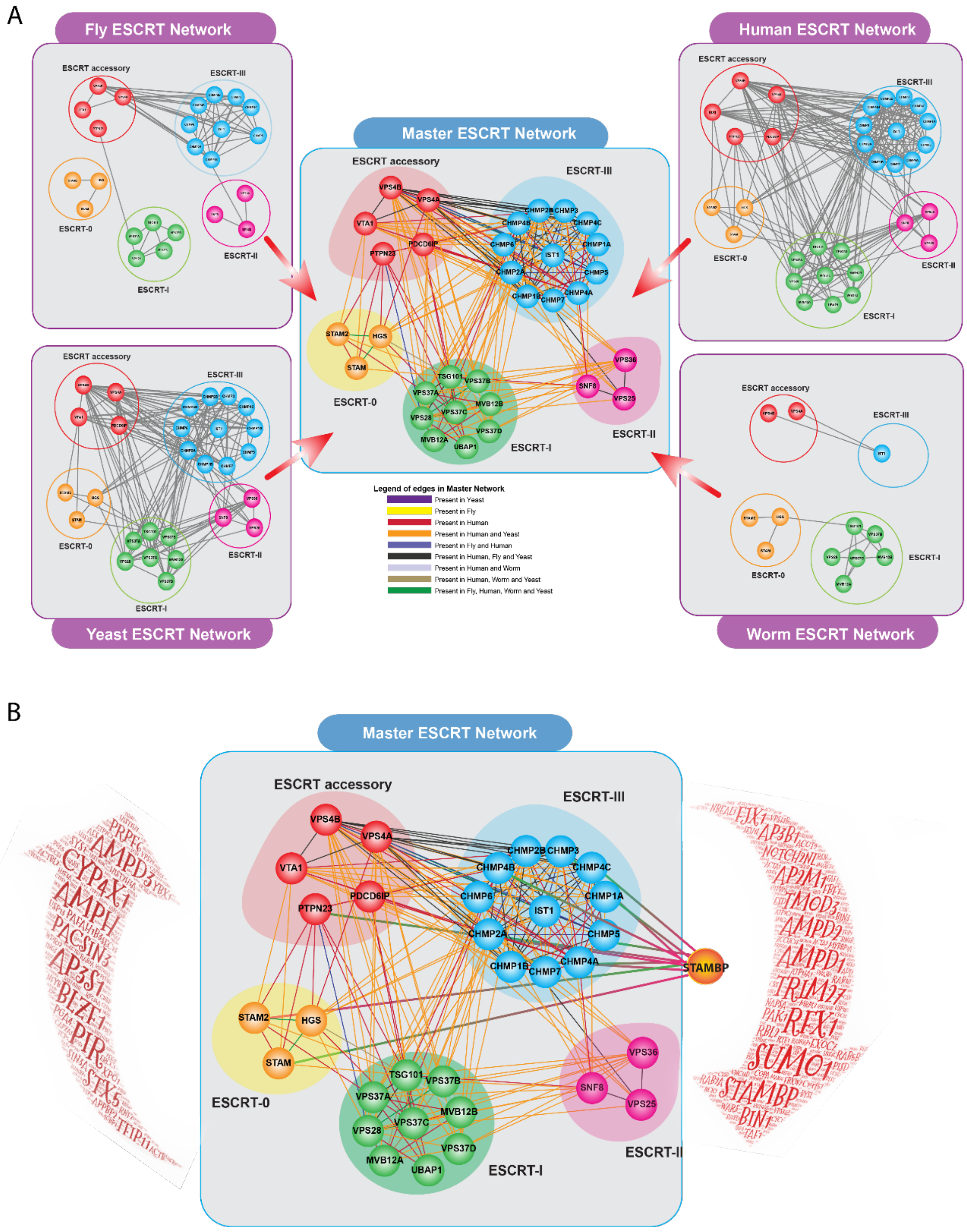


Figure 6-5: Master ESCRT and Test Networks. (A) Organism-specific ESCRT PPI networks are combined to form the master ESCRT network (B) STAMBP is added to the Master ESCRT Network, G_m to form a test network for STAMBP. The arrows depict the iterative process of adding one neighbouring protein to the master network to form a test network.

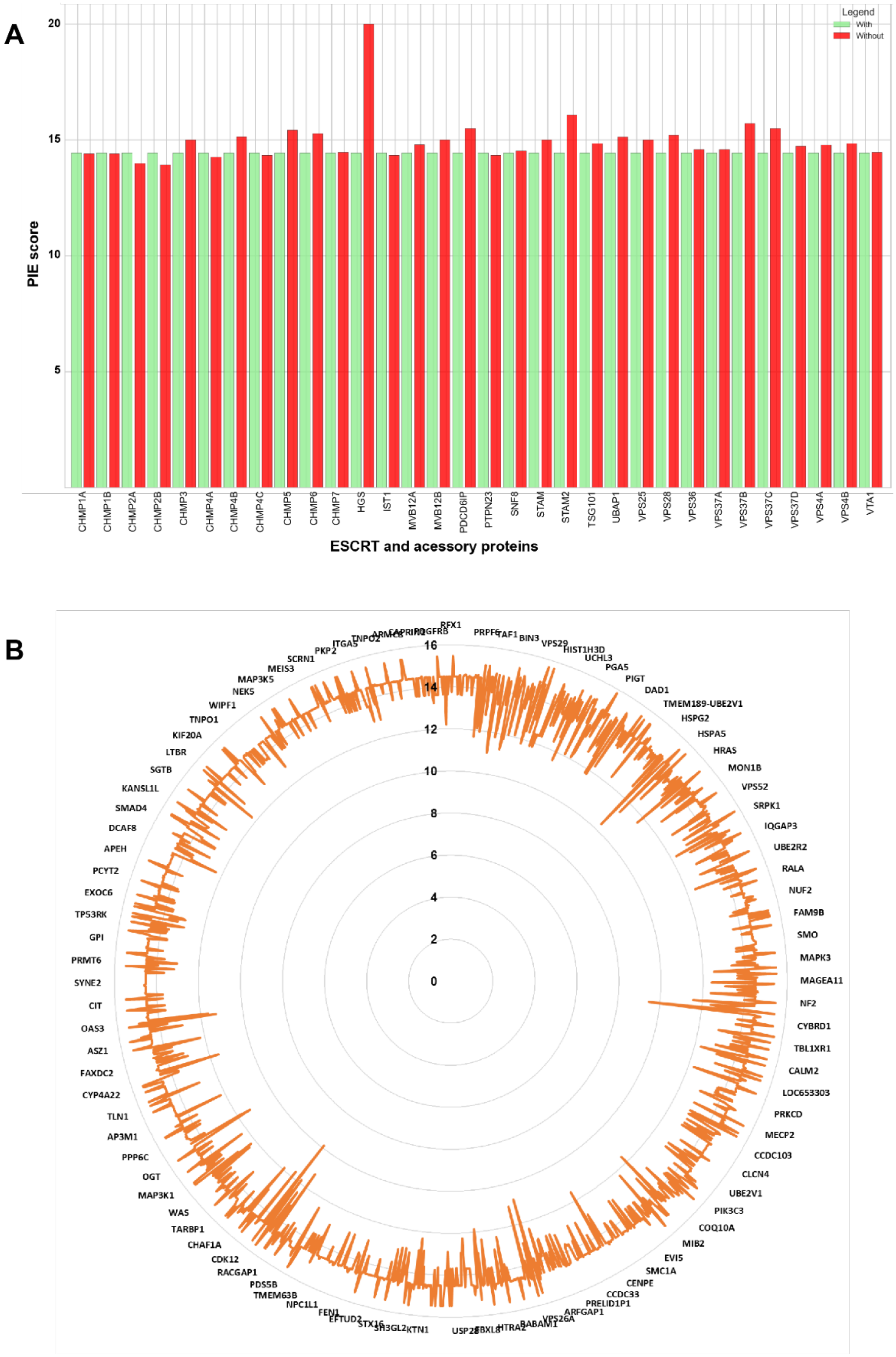


Figure 6-6: Bar plot and radar plots of the distribution of pie scores. A) The overall PIE score for the ESCRT machinery is calculated (shown by green bars) and then for each

ESCRT protein, removed from the network and the remaining ESCRT machinery's physical coherence is calculated (shown by red bars). The resulting PIE scores are compared with the overall network PIE for changes in the physical coherence of the network when each protein is removed from the network. B) Shows the PIE scores for the ESCRT neighbours, spikes pointing inwards represent neighbours that had a PIE score less than the average of 14.2 while spikes pointing outwards had a PIE score higher than the average.

6.4.4. Prediction of NEDD4 and STAMBP as novel regulators of exosome biogenesis

The physical coherence of the master network when a neighbouring protein is added to the network is calculated for each of the 1,800 neighbours and ranked. Figure 6-6 (B) shows the distribution of PIE scores among ESCRT neighbouring proteins. The average PIE score among the neighbours was 14.2 which was slightly less than the ESCRT's 14.4 which we had elected to be a reference point for any physical coherence changes in the master network as discussed above. Neighbouring proteins that changed the physical coherence of the master network by ± 0.5 than the 14.4 reference point were shortlisted as potential candidates for further analysis.

From the shortlisted candidates above, we performed a cellular component enrichment analysis using FunRich [242] and selected only those proteins that enriched for the terms cytoplasm and/or cytosol. In addition to physical coherence, we also used GO function semantic similarity scores to validate the physical coherence scores and narrow down the selected proteins to those that are indeed functionally related to the ESCRT machinery and are biologically relevant to the process of exosome biogenesis. Using a semantic similarity threshold of an average of 0.6, we selected all the proteins that had semantic scores above the set threshold for semantic similarity as predicted potential candidates were expected to regulate exosome biogenesis. Figure 6-7 provides a summary of the relationship between physical coherence and the GO function semantic scores as well as the node degrees of the neighbour proteins with the master network.

From the over 1,800 neighbouring proteins identified above, the final list of shortlisted candidates consisted of 193 (shown in supplementary table 6.1) ESCRT neighbouring proteins as potential candidates likely to regulate exosome biogenesis. Their interactions with the rest of the ESCRT machinery members are mapped, as shown in Figure 6-8. Further functional enrichment analysis was performed on the 193 candidates using FunRich and they were found to be highly enriched for ubiquitin-specific protease activity, chaperone activity, GTPase activity, transporter activity and receptor signalling complex scaffolding activity, as shown in Figure 6-9 (A) and (B).

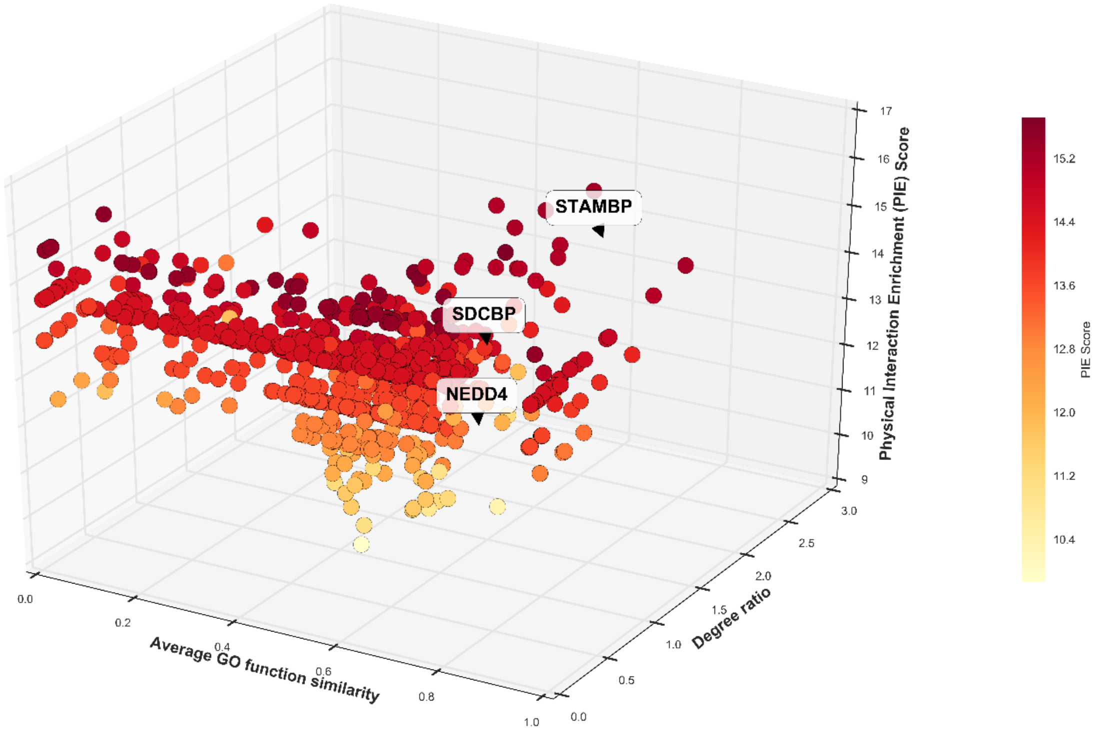


Figure 6-7: 3D scatter plot of PIE vs GO similarity and degree ratio. The physical coherence (PIE) scores for the shortlisted candidates are compared against the GO function semantic similarity and the degree ratio. Based on the literature, the proteins NEDD4 and STAMBP are selected as novel regulators of exosome biogenesis. SDCBP is known to be a regulator of exosome biogenesis and was therefore selected as a control for further laboratory experiments to validate the findings

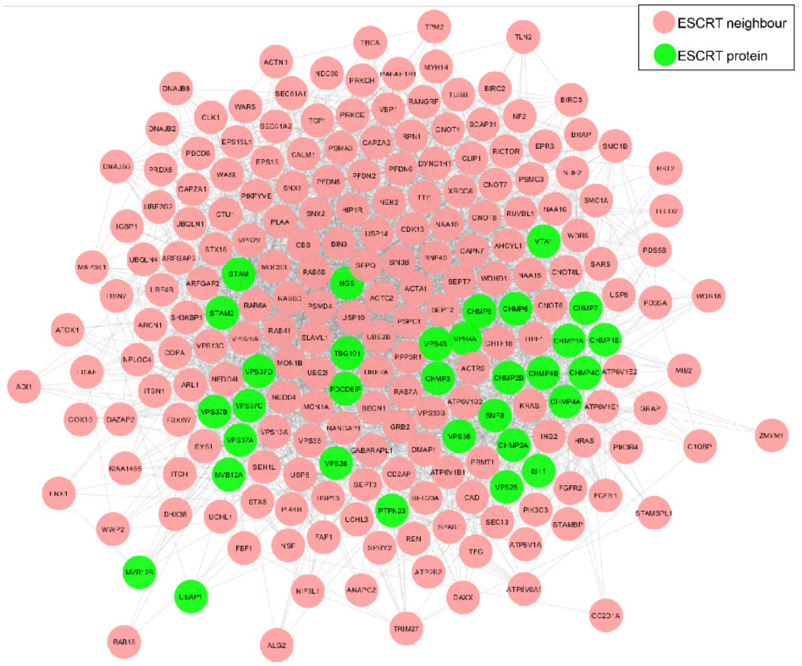


Figure 6-8: ESCRT proteins and shortlisted neighbours. Shortlisted proteins are mapped to their corresponding ESCRT proteins to form a network.

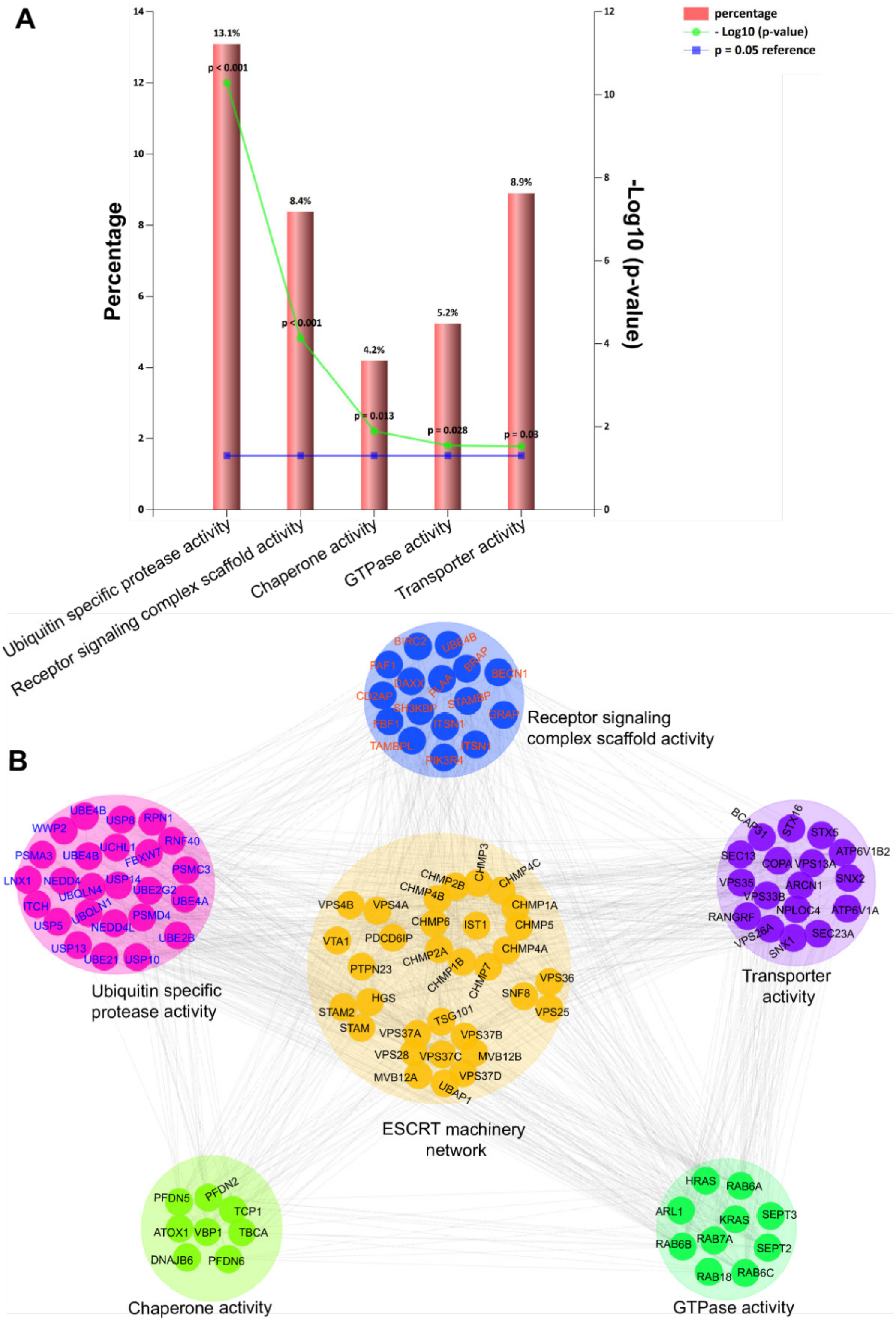


Figure 6-9: Molecular function enrichment (A) the shortlisted proteins are enriched for their molecular functions using FunRich [242] (B) the enriched proteins are then mapped to their corresponding functionalities in a network using Cytoscape [105]

To further validate our list of predicted proteins, we also performed a literature search through PubMed for each of the 193 proteins against the terms ‘viral budding’ and/or ‘exosomes’. The results were manually curated to shortlist potential candidates. Of the 193 predicted proteins, 28 were found to be associated with the term ‘exosomes’, and ten were found to be associated with the term ‘viral budding’ while only NEDD4 was found to be associated with both terms. Figure 6-10 provides a summary of the proteins that were found to have an association with the terms.

Based on the results of the literature search and the PIE scores, we selected two proteins, STAMBP and NEDD4, for experimental validation. STAMBP had a PIE score of 15.1 which was more than the reference PIE score of 14.4 for the ESCRT network. The results imply that the addition of STAMBP to the ESCRT network increased its physical cohesiveness. NEDD4, on the other hand, had a PIE score of 12.4 which in this instance was lower than the reference ESCRT PIE score of 14.4 and therefore meant that the addition of NEDD4 to the ESCRT network decreased the physical cohesiveness of the complex. In addition to the two proteins, we also selected SDCBP as a control as it has been shown previously to regulate exosome biogenesis [324]. Further analysis using CRISPR-Cas9 based knockout cells of NEDD4 and STAMBP confirmed their active role in exosome biogenesis.

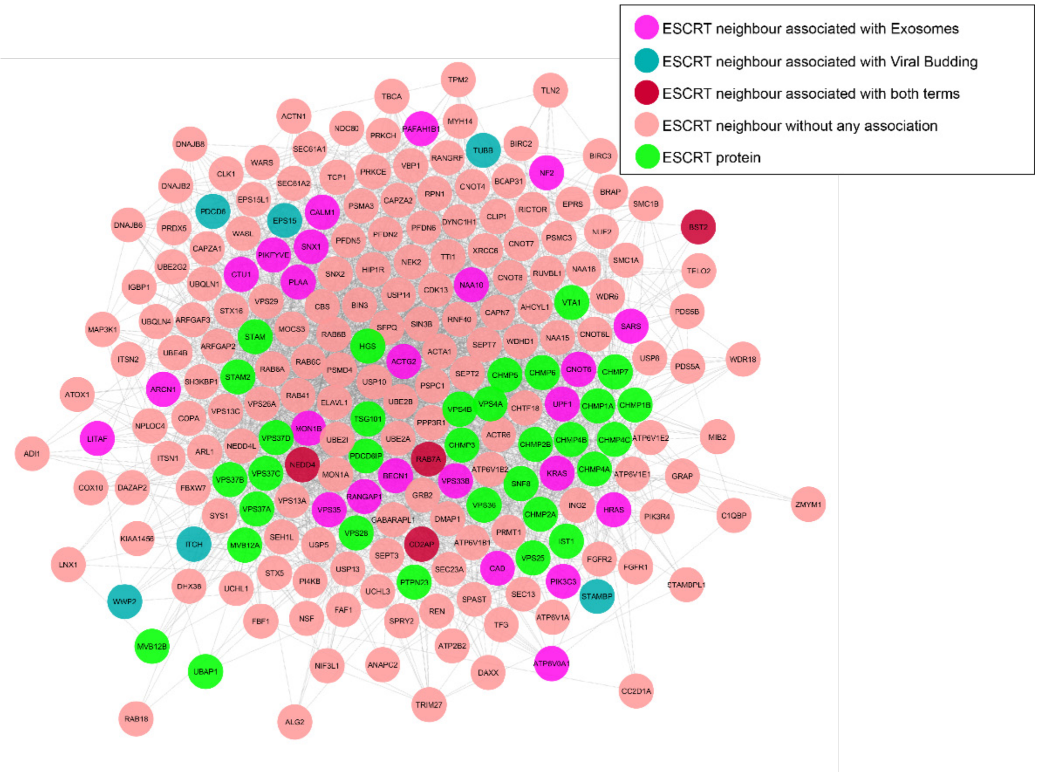


Figure 6-10: ESCRT neighbours associated with exosomes from PubMed. The 193 shortlisted candidates are searched through PubMed for association with the terms exosomes and/or viral budding. The colour coding indicates the proteins that matched or did not match the terms

6.5. Conclusion

As the role of exosomes in both pathological and physiological conditions continues to be unravelled through ongoing research, it is also vital that we understand the mechanisms that constitute exosome biogenesis. It has been shown that the ESCRT machinery is a crucial complex that is known to play a significant role in exosome biogenesis, however, very little is known about the mechanism behind exosome biogenesis. Therefore, in this chapter, we embarked on identifying potential proteins that can regulate exosome biogenesis through the ESCRT machinery. Using computational tools and network theory, we predicted that STAMBP and NEDD4 are potential novel regulators of exosome biogenesis. We found that STAMBP increased the physical coherence of the ESCRT machinery network when its interactions were added to the network while NEDD4 reduced the physical coherence of the ESCRT machinery network. Also, both had significant GO function semantic similarity with the ESCRT proteins. A search through the literature in PubMed further indicated that both STAMBP and NEDD4 had been associated with exosomes or viral budding. The results were validated using CRISPR-Cas9-based knockout cells of NEDD4 and STAMBP, and their active roles in exosome biogenesis were confirmed.

Chapter 7

General discussion

This thesis aimed to develop bioinformatics tools and resources that utilises network theory methods for the analysis of cancer related "omics" data. This was achieved by further dividing the main aim into three other sub-aims which consisted of developing an integrated repository for CRC related "omics" data, identification of essential genes in CRC using network theory and machine learning based techniques and prediction of novel proteins that regulate exosome biogenesis. This chapter discusses the results and findings of this thesis together with some of the significant limitations, future directions and the relevance of our findings to systems biology, and cancer research.

7.1 Colorectal Cancer Atlas: An integrative resource for genomic and proteomic annotations from colorectal cancer cell lines and tissues

The Colorectal Cancer Atlas is an integrated web-based platform offering CRC researchers the opportunity to analyse genomics and proteomics information for genes that have been associated with colorectal cancer in the literature. In recent years, there has been exponential growth in the amounts of heterogeneous omics data obtained from CRC patient samples and cell lines, thanks in part to the introduction of new high-throughput data collection techniques, such as NGS and MS. In Chapter 2, after an extensive literature review, we found that there was a lack of a database resource that could integrate such types of heterogeneous data, and that also enabled the analysis of multidimensional data sets, explicitly relating to CRC. We, therefore, collated this data from the literature as well as database resources and in Chapter 3 [11], I developed The Colorectal Cancer Atlas (http://www.colonatlas.org) which catalogues such data as sequence variants along with quantitative and non-quantitative proteomics data. With this resource, researchers can analyse data in the context of signaling pathways, protein-protein interactions, gene ontology terms, protein domains and post-translational modifications.

Presently, the Colorectal Cancer Atlas comprises data for >13 711 CRC tissues and >179 CRC cell lines. This data includes 62 251 protein identifications, >8.3 million MS/MS spectra, >18 410 genes with sequence variations (404 278 entries) and 351 pathways with sequence variants. This data is continuously updated as new data becomes available. Since its launch, the Colorectal Cancer Atlas has had more than 195,000 page views and has been used by over 21,000 unique users from countries such as the United States of America, United Kingdom, China, Germany and South Korea. This, therefore, highlights the significance of such a resource to the research community.

After the publication of the paper describing this resource in 2015, a similar database resource has since been developed, CoReCG [325]. The resource provides a catalogue of genomic data related to CRC. While CoReCG provides a platform with similar functionality to ours, unlike our tool which integrates both genomic and proteomic data, CoReCG only focuses on genomic-related data. Furthermore, our resource provides a comprehensive collection of literature about CRC genes. Overall, the data collated into the Colorectal Cancer Atlas formed the basis for further analysis in Chapters 4 and 5.

7.2 Perturbation of protein-protein interaction network based on APC mutations in Colorectal Cancer

The identification of essential genes in the tumorigenesis, proliferation and metastasis of cancer remains one of the significant challenges in cancer research due to the heterogeneous nature of cancer. Mutations in APC have been shown to be essential in the tumorigenesis of CRC [243]. It is against this background that in Chapter 4, we set out to understand the topological changes that take place in PPI networks when APC is mutated and therefore, attempted to identify genes that are essential for the proliferation and viability of tumour cells in CRC when APC is mutated. I used the data that we collated into the Colorectal Cancer Atlas and utilised other heterogeneous omics datasets deposited in various online repositories.

In this chapter, I used the concept of the ‘small-world’ property [244] of PPI networks, whereby mutational changes in one gene are cascaded beyond its direct interacting partners. I, therefore, sought to define a method that could integrate the various omics data, while at the same time quantifies the dynamic changes that take place in a PPI network due to mutational and differential gene expression changes. In Chapter 4, I developed a novel method, local area connectivity (LAC), that perturbs a PPI network by globally cascading APC’s mutation information through a PPI network and quantifies the topological changes arising from the perturbation. The method uses node degree to calculate the connectivity of a node but in the process, penalises interacting partners that are either downregulated because of a mutation in APC or are known to be cancer related and are either themselves mutated or not.

Using this method, I predicted over 1600 genes as having significant topological changes in their local connectivity which included already known candidates as well as new potential candidates. Enrichment analysis of the predicted results showed that they were significantly enriched for such biological pathways as cell growth, immune response, cell communication and signal transduction which are all considered to be important in the tumorigenesis, proliferation and metastasis of cancer [16].I found that genes whose average LAC samples were highly enriched for signal transduction and cell communication while those with low LAC scores were highly enriched for immune response and cell growth/maintenance. Here, we hypothesised that when APC is mutated, genes which are involved in the signal transduction and cell communication pathways become highly interconnected because of APC mutations and consequently, become irresponsive to new signals as the signalling processes are never terminated. This view is supported by several previous research studies which documented the roles of these pathways in cancer [18, 245]. On the contrary, when APC is mutated, genes which are involved in immune response and cell growth/maintenance are less well connected. As such, their respective pathways may become inactivated and which, in turn, may enhance the proliferation of cancer cells as the mechanisms responsible for controlling cell growth and/or maintenance or immune response are lost [246]. However, the high enrichment for immune response in this case can also be explained by the fact that gene expression profiling was conducted in cancer cell line.

To validate the predicted results, I used the Achilles dataset and found several of the predicted genes to be essential in the proliferation of CRC cancer cell lines. Among the identified genes included direct APC interactors such as AXIN2, CTNNB1, DKK3, KRT23, KRT5 and NOSTRIN which were also found to be underexpressed in the SW480+APC cell line. The roles of AXIN2 and CTNNB1 in CRC have been well documented in the literature [33, 243, 247, 248]. I, therefore, used these as references to understand the roles of the other genes when analysing the Achilles data set and as a result, we found that DKK3 and KRT23 are essential in cell proliferation in mutant APC cell lines. This observation was confirmed further by results from the gene expression profiling of

TCGA patient samples which showed that the median expressions of DKK3 and KRT23 in mutant APC samples are higher than that in wild-type APC samples. This observation is supported by previous research [249] where it has been shown that DKK3 is overexpressed in CRC and Birkenkamp-Demtröder, et al. [250] also showed that by knocking down KRT23 in CRC samples, cell proliferation is reduced. NOSTRIN and KRT5 did not show notable differences in their Achilles scores.

In addition to APC interactors, I also profiled genes that are not direct interactors of APC but had significant LAC scores and were also found to be essential for the viability of CRC cancer cell lines. Notable among these were STAT3[251] and TSG101[252] which have been implicated in CRC while others included APOBEC3G, ASL, CDKN1A, CTSH, DDIT4, DYNC1H1, ELMO3, EVPL, FUBP1, GNA11, GPSM1, NFKBIA, PRSS2, PSMA1, SUPT5H, TMOD1, TRAF1, TRIP13, TUBA1B and USP39. A search on PubMed for these genes revealed that they had been found to play a role in CRC or some other form of cancer. For instance, APOBEC3G, a gene that codes for the protein apolipoprotein B mRNA-editing enzyme catalytic polypeptide-like 3G has been found to be associated with poor prognosis in CRC patients [253]. ASL, a gene that codes for arginosuccinate lyase, has also been shown to dysregulate nitric oxide (an essential mediator in the tumorigenesis of various cancers) which, in turn, inhibits the proliferation of cancers such as liver and breast cancer [254]. The remaining genes have also been implicated in functions ranging from tumorigenesis to metastasis [255-259]. While the roles of some of the predicted genes have been documented, there is still a need to understand their roles in the progression of CRC when APC is mutated. In this chapter,I have shown that when APC is mutated, these genes undergo significant changes in their interactions within a PPI network and therefore are likely to play an essential role in CRC. The findings in this chapter have significant implications towards unravelling the role of passenger genes in CRC research. Nonetheless, caution should be taken when applying LAC in contexts where the effect of a mutation on the cell phenotype is known in advance as well as those where certain PPIs are known to be either active or not.

7.3 Integration of heterogeneous 'omics' data using semi-supervised network labelling to identify essential genes in colorectal cancer.

In Chapter 5, building on the work discussed in Chapter 4, I applied a semi-supervised machine learning technique to infer genes that are likely to be affected when the PPI network topology changes due to genomic changes such as mutations and gene expression variations. Unlike in Chapter 4 where I focused on the role of APC mutations on the topology of the PPI network, in Chapter 5, I sought to understand the collective effect of all known gene mutations across a range of samples on the topology of the PPI network. In Chapter 2, literature review showed that many of the network analysis methods that are applied in identifying essential genes are often based on the concept of "guilty by association" (GBA). GBA methods are based on the premise that proteins within a PPI network that interact with proteins that are known to be associated with a given disease are themselves considered to be associated with that disease. While such methods have, over the years, been useful in identifying essential genes, recent research has shown that attempts to apply the same principle to an entire PPI network have yielded varying results. As such, in Chapter 5, I applied network propagation, a semi-supervised labelling technique that allows genomic information or signal changes in one protein to be cascaded beyond its direct interacting partners.

Although the technique is not new to this thesis, I used it to validate some already known pathways, as well as genes that have been implicated in CRC. I further used the technique to predict novel genes that are affected by the genomic changes in other genes, even when they are not directly interacting partners of genes that had undergone genomic changes. Examples of the genes identified include ASXL1, SMAD4, RALY, DIDO1 and CCAR2. Several publications have applied this method to the identification of essential genes in disease [164, 166, 174], however none has conducted such a study specific to CRC as ourselves. Nonetheless, one of the challenges in computational predictions is that they are usually prone to false positives [326], thus the need for the experimental validation of the predicted results. In this chapter, the roles of the identified proteins in CRC were not validated through experimentation, and as such, further experiments will therefore be required. This chapter nevertheless demonstrated that by integrating heterogeneous data types, overlaying this information on a PPI network and propagating the information beyond immediate interacting partners, I could infer the effect of such changes on the entire global PPI network and therefore was able to quantify the distance between the two proteins. In addition, I further identified proteins that have been implicated in other types of cancers which are likely to have some implications in CRC.

7.4 Physical coherence and network analysis to identify novel regulators of exosome biogenesis.

In Chapter 6, we identified NEDD4 and STAMBP proteins as novel regulators of exosome biogenesis through the ESCRT pathway. These were further verified by wet laboratory experimentation and analysis using CRISPR-Cas9-based knockout cells of NEDD4 and STAMBP which confirmed their active role in exosome biogenesis. The growing importance of exosomes in biomedical research, as well as the lack of a well-established mechanism behind exosome biogenesis, necessitated this study. The roles carried out by exosomes in both physiological and pathological states have, over the years, been uncovered and have ranged from intercellular communication [186, 191], disease [197], to being potent drug delivery vehicles [201, 305].

Nevertheless, an understanding of the mechanism by which exosomes are formed, packaged and released to the extracellular environment remains elusive. Regardless of this, several pathways have been identified as potential mechanisms behind exosome biogenesis, such as the endosomal sorting complex for transport (ESCRT) pathway and ESCRT independent pathways consisting of tetraspanins, Rab GTPases and lipids. In Chapter 4, we therefore used computational methods and network theory to the analysis of the ESCRT PPI network to uncover the topological changes in the ESCRT pathway when ESCRT interacting partners are repeatedly added to and removed from the ESCRT PPI network. In Chapter 2, we found that one of the downsides of using PPIs in the inference of essential genes was the fact that PPIs are themselves prone to false discoveries due to experimental errors and literature bias. That is, genes that have been well studied or have been found to be implicated in disease are frequently found to have a high number of interacting partners, as opposed to those that are less studied. Against this background, we incorporated physical coherence [9] into our pipeline to handle the bias in PPI interactions. Using this method, we profiled and computed the physical interaction enrichment scores for over 1800 ESCRT PPI interacting partners, of which 193 were found to change the physical coherence of the ESCRT PPI network significantly. Upon further validation, we found STAMBP and

NEDD4 as novel regulators of exosome biogenesis. To our knowledge, no other study has used computational analysis to implicate these two proteins in exosome biogenesis. The findings in this chapter, therefore, have substantial implications for improving our understanding of the mechanisms behind exosome biogenesis.

7.5 Future directions

The future direction of this study will be the redesign of the Colorectal Cancer Atlas described in chapter 3 to incorporate the method (Local Area Connectivity) developed in Chapter 4 to help users compute the essentiality of genes in the tumour proliferation and viability of CRC cells. This improvement will, in turn, help biomedical researchers who conduct wet laboratory experiments to determine beforehand which genes to focus on. We expect this will, in turn, help lower the costs and time associated with conducting such studies. Furthermore, to enhance the functionality and accessibility of Colorectal Cancer Atlas, the resource will be redesigned so as make it mobile device compatible as well as the provision of a submission page for new CRC related data or modification of existing data by other researches. The inclusion of a submission page will help improve the quality as well as the amount of data that Colorectal Cancer Atlas will be able to support.

Additionally, future works for the work described in chapter 5 include the fine tuning of the model used and incorporating new machine learning techniques such as deep learning using neural networks to the integration of heterogeneous datasets and inference of essential genes in CRC. The genes identified in Chapter 5 will also need validation through the conduction of wet laboratory based experiments.

References

[1] E. F. Petricoin *et al.*, "Mapping molecular networks using proteomics: a vision for patient-tailored combination therapy," *Journal of clinical oncology,* vol. 23, no. 15, pp. 3614-3621, 2005.

[2] N. James, Cancer : A Very Short Introduction, Oxford: Oxford University Press, 2011. [Online]. Available: http://latrobe.eblib.com.au/patron/FullRecord.aspx?p=746766.

[3] WHO. (2017). *Cancer Fact Sheet*. Available: http://www.who.int/mediacentre/factsheets/fs297/en/

WHO2017 - Cancer Fact Sheet.html

[4] WHO. (2015, 27/07/2015). *Cancer*. Available: http://www.who.int/mediacentre/factsheets/fs297/en/

[5] A. C. Society, Cancer: What Causes It, What Doesn't, Chicago: American Cancer Society, 2012. [Online]. Available: http://latrobe.eblib.com.au/patron/FullRecord.aspx?p=882986.

[6] H. Y. Chuang, E. Lee, Y. T. Liu, D. Lee, and T. Ideker, "Network-based classification of breast cancer metastasis," *Molecular systems biology,* vol. 3, no. 1, 2007.

[7] E. Zuckerkandl and L. Pauling, "Molecular disease, evolution and genetic heterogeneity," 1962.

[8] A.-L. Barabási, N. Gulbahce, and J. Loscalzo, "Network medicine: a network-based approach to human disease," *Nature Reviews Genetics,* vol. 12, no. 1, pp. 56-68, 2011.

[9] I. E. Sama and M. A. Huynen, "Measuring the physical cohesiveness of proteins using physical interaction enrichment," *Bioinformatics,* vol. 26, no. 21, pp. 2737-2743, 08/26

06/10/received

08/12/revised

08/13/accepted 2010.

[10] T. Sevimoglu and K. Y. Arga, "The role of protein interaction networks in systems biomedicine," *Computational and Structural Biotechnology Journal,* vol. 11, no. 18, pp. 22-27, 09/03 2014.

[11] D. Chisanga *et al.*, "Colorectal cancer atlas: An integrative resource for genomic and proteomic annotations from colorectal cancer cell lines and tissues," *Nucleic Acids Research,* vol. 44, no. D1, pp. D969-D974, January 4, 2016 2016.

[12] G. S. Cowley *et al.*, "Parallel genome-scale loss of function screens in 216 cancer cell lines for the identification of context-specific genetic dependencies," Data Descriptor vol. 1, p. 140035, 09/30/online 2014.

[13] World Cancer Research Fund International. (2017, 20 May). *Colorectal Cancer Statistics*. Available: http://www.wcrf.org/int/cancer-facts-figures/data-specific-cancers/colorectal-cancer-statistics

[14] M. Arnold, M. S. Sierra, M. Laversanne, I. Soerjomataram, A. Jemal, and F. Bray, "Global patterns and trends in colorectal cancer incidence and mortality," *Gut,* pp. gutjnl-2015-310912, 2016.

[15] D. Hanahan and R. A. Weinberg, "The hallmarks of cancer," *cell,* vol. 100, no. 1, pp. 57-70, 2000.

[16] D. Hanahan and Robert A. Weinberg, "Hallmarks of Cancer: The Next Generation," *Cell,* vol. 144, no. 5, pp. 646-674, 3/4/ 2011.

[17] R. Fodde, "The APC gene in colorectal cancer," *European Journal of Cancer,* vol. 38, no. 7, pp. 867-871, 2002/05/01/ 2002.

[18] R. Fodde, R. Smits, and H. Clevers, "APC, Signal transduction and genetic instability in colorectal cancer," *Nat Rev Cancer,* 10.1038/35094067 vol. 1, no. 1, pp. 55-67, 10//print 2001.

[19] K. Bardhan and K. Liu, "Epigenetics and Colorectal Cancer Pathogenesis," *Cancers,* vol. 5, no. 2, p. 676, 2013.

[20] M. F. Müller, A. E. K. Ibrahim, and M. J. Arends, "Molecular pathological classification of colorectal cancer," *Virchows Archiv,* vol. 469, pp. 125-134, 06/20

03/30/received

05/04/revised

05/09/accepted 2016.

[21] E. R. Fearon, "Molecular Genetics of Colorectal Cancer," *Annual Review of Pathology: Mechanisms of Disease,* vol. 6, no. 1, pp. 479-507, 2011.

[22] M. S. Pino and D. C. Chung, "The chromosomal instability pathway in colon cancer," *Gastroenterology,* vol. 138, no. 6, pp. 2059-2072, 2010.

[23] C. Lengauer, K. W. Kinzler, and B. Vogelstein, "Genetic instabilities in human cancers," *Nature,* 10.1038/25292 vol. 396, no. 6712, pp. 643-649, 12/17/print 1998.

[24] E. Fearon and B. Vogelstein, "A genetic model for colorectal tumorigenesis," *Cell,* vol. 61, 1990.

[25] J. Schneikert and J. Behrens, "The canonical Wnt signalling pathway and its APC partner in colon cancer development," *Gut,* vol. 56, no. 3, pp. 417-425, 2007.

[26] Y. Samuels and T. Waldman, "Oncogenic mutations of PIK3CA in human cancers," (in eng), *Curr Top Microbiol Immunol,* vol. 347, pp. 21-41, 2010.

[27] L. D. Wood *et al.*, "The Genomic Landscapes of Human Breast and Colorectal Cancers," *Science,* vol. 318, no. 5853, pp. 1108-1113, 2007.

[28] TCGAN, "Comprehensive molecular characterization of human colon and rectal cancer," *Nature,* 10.1038/nature11252 vol. 487, no. 7407, pp. 330-337, 07/19/print 2012.

[29] R. J. Leary *et al.*, "Integrated analysis of homozygous deletions, focal amplifications, and sequence alterations in breast and colorectal cancers," *Proceedings of the National Academy of Sciences,* vol. 105, no. 42, pp. 16224-16229, October 21, 2008 2008.

[30] A. J. Rowan *et al.*, "APC mutations in sporadic colorectal tumors: A mutational "hotspot" and interdependence of the "two hits"," *Proceedings of the National Academy of Sciences of the United States of America,* vol. 97, no. 7, pp. 3352-3357, 12/22/accepted 2000.

[31] K. M. Haigis *et al.*, "Differential effects of oncogenic K-Ras and N-Ras on proliferation, differentiation and tumor progression in the colon," (in eng), *Nat Genet,* vol. 40, no. 5, pp. 600-8, May 2008.

[32] L. N. Kwong and W. F. Dove, "APC and its modifiers in colon cancer," *Advances in experimental medicine and biology,* vol. 656, pp. 85-106, 2009.

[33] T. Nakamura *et al.*, "Axin, an inhibitor of the Wnt signalling pathway, interacts with β-catenin, GSK-3β and APC and reduces the β-catenin level," *Genes to Cells,* vol. 3, no. 6, pp. 395-403, 1998.

[34] J. Groden *et al.*, "Identification and characterization of the familial adenomatous polyposis coli gene," (in eng), *Cell,* vol. 66, no. 3, pp. 589-600, Aug 09 1991.

[35] E. M. Schatoff, B. I. Leach, and L. E. Dow, "WNT Signaling and Colorectal Cancer," *Current Colorectal Cancer Reports,* journal article vol. 13, no. 2, pp. 101-110, April 01 2017.

[36] L. Zhang and J. W. Shay, "Multiple Roles of APC and its Therapeutic Implications in Colorectal Cancer," *JNCI: Journal of the National Cancer Institute,* vol. 109, no. 8, pp. djw332-djw332, 2017.

[37] M. Vidal, M. E. Cusick, and A.-L. Barabasi, "Interactome networks and human disease," *Cell,* vol. 144, no. 6, pp. 986-998, 2011.

[38] A. Blais and B. D. Dynlacht, "Constructing transcriptional regulatory networks," *Genes & development,* vol. 19, no. 13, pp. 1499-1511, 2005.

[39] C. Zhu *et al.*, "High-resolution DNA-binding specificity analysis of yeast transcription factors," *Genome research,* vol. 19, no. 4, pp. 556-566, 2009.

[40] K.-I. Goh, M. E. Cusick, D. Valle, B. Childs, M. Vidal, and A.-L. Barabási, "The human disease network," *Proceedings of the National Academy of Sciences,* vol. 104, no. 21, pp. 8685-8690, 2007.

[41] J. De Las Rivas and C. Fontanillo, "Protein–protein interaction networks: unraveling the wiring of molecular machines within the cell," *Briefings in functional genomics,* vol. 11, no. 6, pp. 489-496, 2012.

[42] G. Kar, A. Gursoy, and O. Keskin, "Human Cancer Protein-Protein Interaction Network: A Structural Perspective," *PLoS Comput Biol,* vol. 5, no. 12, p. e1000601, 2009.

[43] M. G. Kann, "Protein interactions and disease: computational approaches to uncover the etiology of diseases," *Briefings in bioinformatics,* vol. 8, no. 5, pp. 333-346, 2007.

[44] P. F. Jonsson and P. A. Bates, "Global topological features of cancer proteins in the human interactome," *Bioinformatics,* vol. 22, no. 18, pp. 2291-2297, 2006.

[45] M. A. Huynen, B. Snel, C. v. Mering, and P. Bork, "Function prediction and protein networks," *Current opinion in cell biology,* vol. 15, no. 2, pp. 191-198, 2003.

[46] P. Jancura and E. Marchiori, "Dividing protein interaction networks for modular network comparative analysis," *Pattern Recognition Letters,* vol. 31, no. 14, pp. 2083-2096, 2010.

[47] J. De Las Rivas and C. Fontanillo, "Protein–Protein Interactions Essentials: Key Concepts to Building and Analyzing Interactome Networks," *PLOS Computational Biology,* vol. 6, no. 6, p. e1000807, 2010.

[48] R. L. Finley and R. Brent, "Interaction mating reveals binary and ternary connections between Drosophila cell cycle regulators," *Proceedings of the National Academy of Sciences,* vol. 91, no. 26, pp. 12980-12984, 1994.

[49] P. L. Bartel, J. A. Roecklein, D. SenGupta, and S. Fields, "A protein linkage map of Escherichia coli bacteriophage T7," *Nature genetics,* vol. 12, no. 1, pp. 72-77, 1996.

[50] M. Fromont-Racine, J.-C. Rain, and P. Legrain, "Toward a functional analysis of the yeast genome through exhaustive two-hybrid screens," *Nature genetics,* vol. 16, no. 3, pp. 277-282, 1997.

[51] M. Vidal, R. K. Brachmann, A. Fattaey, E. Harlow, and J. D. Boeke, "Reverse two-hybrid and one-hybrid systems to detect dissociation of protein-protein and DNA-protein interactions," *Proceedings of the National Academy of Sciences,* vol. 93, no. 19, pp. 10315-10320, 1996.

[52] T. Gandhi *et al.*, "Analysis of the human protein interactome and comparison with yeast, worm and fly interaction datasets," *Nature genetics,* vol. 38, no. 3, pp. 285-293, 2006.

[53] P. M. Roberts, "Mining literature for systems biology," *Briefings in bioinformatics,* vol. 7, no. 4, pp. 399-406, 2006.

[54] A. M. Cohen and W. R. Hersh, "A survey of current work in biomedical text mining," *Briefings in bioinformatics,* vol. 6, no. 1, pp. 57-71, 2005.

[55] R. A. Erhardt, R. Schneider, and C. Blaschke, "Status of text-mining techniques applied to biomedical text," *Drug discovery today,* vol. 11, no. 7, pp. 315-325, 2006.

[56] E. M. Marcotte and S. V. Date, "Exploiting big biology: integrating large-scale biological data for function inference," *Briefings in bioinformatics,* vol. 2, no. 4, pp. 363-374, 2001.

[57] L. Skrabanek, H. K. Saini, G. D. Bader, and A. J. Enright, "Computational prediction of protein–protein interactions," *Molecular biotechnology,* vol. 38, no. 1, pp. 1-17, 2008.

[58] R. Hosur, "Structure-based algorithms for protein-protein interaction prediction," Massachusetts Institute of Technology, 2012.

[59] C. Stark, B.-J. Breitkreutz, T. Reguly, L. Boucher, A. Breitkreutz, and M. Tyers, "BioGRID: a general repository for interaction datasets," *Nucleic acids research,* vol. 34, no. suppl 1, pp. D535-D539, Jan 01 2006.

[60] H. Zhou and E. Jakobsson, "Predicting Protein-Protein Interaction by the Mirrortree Method: Possibilities and Limitations," *PloS one,* vol. 8, no. 12, p. e81100, 2013.

[61] E. M. Phizicky and S. Fields, "Protein-protein interactions: methods for detection and analysis," *Microbiological Reviews,* vol. 59, no. 1, pp. 94-123, March 1, 1995 1995.

[62] S. Fields and R. Sternglanz, "The two-hybrid system: an assay for protein-protein interactions," *Trends in Genetics,* vol. 10, no. 8, pp. 286-292, 1994.

[63] R. Aebersold and M. Mann, "Mass spectrometry-based proteomics," *Nature,* vol. 422, no. 6928, pp. 198-207, Mar 13 2003.

[64] J. R. Yates, "Mass spectral analysis in proteomics," *Annu. Rev. Biophys. Biomol. Struct.,* vol. 33, pp. 297-316, 2004.

[65] J. R. Yates, C. I. Ruse, and A. Nakorchevsky, "Proteomics by mass spectrometry: approaches, advances, and applications," *Annual review of biomedical engineering,* vol. 11, pp. 49-79, 2009.

[66] A. Bensimon, A. J. Heck, and R. Aebersold, "Mass spectrometry-based proteomics and network biology," *Annual review of biochemistry,* vol. 81, pp. 379-405, 2012.

[67] R. M. Ewing *et al.*, "Large-scale mapping of human protein–protein interactions by mass spectrometry," *Molecular systems biology,* vol. 3, no. 1, 2007.

[68] O. Puig *et al.*, "The Tandem Affinity Purification (TAP) Method: A General Procedure of Protein Complex Purification," *Methods,* vol. 24, no. 3, pp. 218-229, 7// 2001.

[69] M. F. Templin, D. Stoll, M. Schrenk, P. C. Traub, C. F. Vöhringer, and T. O. Joos, "Protein microarray technology," *Drug Discovery Today,* vol. 7, no. 15, pp. 815-822, 8/1/ 2002.

[70] M. Barrios-Rodiles *et al.*, "High-Throughput Mapping of a Dynamic Signaling Network in Mammalian Cells," *Science,* vol. 307, no. 5715, pp. 1621-1625, March 11, 2005 2005.

[71] M. E. Cusick *et al.*, "Literature-curated protein interaction datasets," *Nature methods,* vol. 6, no. 1, pp. 39-46, 2009.

[72] A. L. Turinsky, S. Razick, B. Turner, I. M. Donaldson, and S. J. Wodak, "Literature curation of protein interactions: measuring agreement across major public databases," *Database: the journal of biological databases and curation,* vol. 2010, 2010.

[73] D. Plewczyński and K. Ginalski, "The interactome: predicting the protein-protein interactions in cells," *Cellular & molecular biology letters,* vol. 14, no. 1, pp. 1-22, 2009.

[74] M. W. Gonzalez and M. G. Kann, "Chapter 4: Protein Interactions and Disease," *PLoS Comput Biol,* vol. 8, no. 12, p. e1002819, 2012.

[75] D. Botstein and N. Risch, "Discovering genotypes underlying human phenotypes: past successes for mendelian disease, future approaches for complex disease," *Nature genetics,* vol. 33, pp. 228-237, 2003.

[76] T. Ideker and R. Sharan, "Protein networks in disease," *Genome research,* vol. 18, no. 4, pp. 644-652, 2008.

[77] P. Hallock and M. A. Thomas, "Integrating the Alzheimer's disease proteome and transcriptome: a comprehensive network model of a complex disease," *Omics: a journal of integrative biology,* vol. 16, no. 1-2, pp. 37-49, 2012.

[78] T. Ideker, O. Ozier, B. Schwikowski, and A. F. Siegel, "Discovering regulatory and signalling circuits in molecular interaction networks," *Bioinformatics,* vol. 18, no. suppl 1, pp. S233-S240, 2002.

[79] S. Wachi, K. Yoneda, and R. Wu, "Interactome-transcriptome analysis reveals the high centrality of genes differentially expressed in lung cancer tissues," *Bioinformatics,* vol. 21, no. 23, pp. 4205-4208, 2005.

[80] J. Lim *et al.*, "A protein–protein interaction network for human inherited ataxias and disorders of Purkinje cell degeneration," *Cell,* vol. 125, no. 4, pp. 801-814, 2006.

[81] J. C. Charlesworth *et al.*, "Transcriptomic epidemiology of smoking: the effect of smoking on gene expression in lymphocytes," *BMC medical genomics,* vol. 3, no. 1, p. 29, 2010.

[82] B.-Q. Li, T. Huang, L. Liu, Y.-D. Cai, and K.-C. Chou, "Identification of Colorectal Cancer Related Genes with mRMR and Shortest Path in Protein-Protein Interaction Network," *PLOS ONE,* vol. 7, no. 4, p. e33393, 2012.

[83] E. J. Rossin *et al.*, "Proteins encoded in genomic regions associated with immune-mediated disease physically interact and suggest underlying biology," *PLoS genetics,* vol. 7, no. 1, p. e1001273, 2011.

[84] A. A. Ivanov, F. R. Khuri, and H. Fu, "Targeting protein–protein interactions as an anticancer strategy," *Trends in Pharmacological Sciences,* vol. 34, no. 7, pp. 393-400, 7// 2013.

[85] O. B. Poirion, X. Zhu, T. Ching, and L. Garmire, "Single-Cell Transcriptomics Bioinformatics and Computational Challenges," (in English), *Frontiers in Genetics,* Mini Review vol. 7, no. 163, 2016-September-21 2016.

[86] D. Chisanga, S. Keerthikumar, and N. Chilamkurti, "Network tools for the analysis of proteomic data," in *Proteome Bioinformatics*, S. Mathivanan, Ed.: Springer, 2017.

[87] S. Mathivanan, "Integrated bioinformatics analysis of the publicly available protein data shows evidence for 96% of the human proteome," *Journal of Proteomics & Bioinformatics,* vol. 2014, no. 7, pp. 041-049, 2014.

[88] M.-S. Kim *et al.*, "A draft map of the human proteome," *Nature,* Article vol. 509, no. 7502, pp. 575-581, 05/29/print 2014.

[89] M. Wilhelm *et al.*, "Mass-spectrometry-based draft of the human proteome," *Nature,* Article vol. 509, no. 7502, pp. 582-587, 05/29/print 2014.

[90] G. A. Pavlopoulos *et al.*, "Using graph theory to analyze biological networks," *BioData Mining,* vol. 4, no. 1, pp. 1-27, 2011// 2011.

[91] T. K. B. Gandhi *et al.*, "Analysis of the human protein interactome and comparison with yeast, worm and fly interaction datasets," *Nat Genet,* 10.1038/ng1747 vol. 38, no. 3, pp. 285-293, 03//print 2006.

[92] S. Mathivanan *et al.*, "An evaluation of human protein-protein interaction data in the public domain," *BMC Bioinformatics,* journal article vol. 7, no. 5, pp. 1-14, 2006.

[93] M. Pathan *et al.*, "FunRich: An open access standalone functional enrichment and interaction network analysis tool," *PROTEOMICS,* vol. 15, no. 15, pp. 2597-2601, 2015.

[94] M. R. Wilkins *et al.*, "From Proteins to Proteomes: Large Scale Protein Identification by Two-Dimensional Electrophoresis and Arnino Acid Analysis," *Nat Biotech,* 10.1038/nbt0196-61 vol. 14, no. 1, pp. 61-65, 01//print 1996.

[95] A. Schmidt, I. Forne, and A. Imhof, "Bioinformatic analysis of proteomics data," *BMC Systems Biology,* vol. 8, no. Suppl 2, pp. S3-S3, 03/13 2014.

[96] J. De Las Rivas and C. Fontanillo, "Protein–Protein Interactions Essentials: Key Concepts to Building and Analyzing Interactome Networks," *PLoS Comput Biol,* vol. 6, no. 6, p. e1000807, 2010.

[97] T. S. Keshava Prasad *et al.*, "Human Protein Reference Database--2009 update," (in eng), *Nucleic Acids Res,* vol. 37, no. Database issue, pp. D767-72, Jan 2009.

[98] L. Licata *et al.*, "MINT, the molecular interaction database: 2012 update," *Nucleic Acids Research,* vol. 40, no. D1, pp. D857-D861, January 1, 2012 2012.

[99] A. Chatr-Aryamontri *et al.*, "The BioGRID interaction database: 2015 update," (in eng), *Nucleic Acids Res,* vol. 43, no. Database issue, pp. D470-8, Jan 2015.

[100] D. Szklarczyk *et al.*, "STRING v10: protein-protein interaction networks, integrated over the tree of life," (in eng), *Nucleic Acids Res,* vol. 43, no. Database issue, pp. D447-52, Jan 2015.

[101] L. Salwinski, C. S. Miller, A. J. Smith, F. K. Pettit, J. U. Bowie, and D. Eisenberg, "The Database of Interacting Proteins: 2004 update," *Nucleic Acids Research,* vol. 32, no. suppl 1, pp. D449-D451, January 1, 2004 2004.

[102] G. D. Bader, D. Betel, and C. W. Hogue, "BIND: the Biomolecular Interaction Network Database," (in eng), *Nucleic Acids Res,* vol. 31, no. 1, pp. 248-50, Jan 1 2003.

[103] S. Orchard *et al.*, "The MIntAct project—IntAct as a common curation platform for 11 molecular interaction databases," *Nucleic Acids Research,* vol. 42, no. D1, pp. D358-D363, January 1, 2014 2014.

[104] A. Chatr-aryamontri *et al.*, "MINT: the Molecular INTeraction database," *Nucleic Acids Research,* vol. 35, no. suppl 1, pp. D572-D574, January 1, 2007 2007.

[105] P. Shannon *et al.*, "Cytoscape: a software environment for integrated models of biomolecular interaction networks," *Genome research,* vol. 13, no. 11, pp. 2498-2504, 2003.

[106] M. Kohl, S. Wiese, and B. Warscheid, "Cytoscape: software for visualization and analysis of biological networks," *Data Mining in Proteomics: From Standards to Applications,* pp. 291-303, 2011.

[107] A.-L. Barabasi and Z. N. Oltvai, "Network biology: understanding the cell's functional organization," *Nat Rev Genet,* 10.1038/nrg1272 vol. 5, no. 2, pp. 101-113, 02//print 2004.

[108] Y. Ho *et al.*, "Systematic identification of protein complexes in Saccharomyces cerevisiae by mass spectrometry," *Nature,* 10.1038/415180a vol. 415, no. 6868, pp. 180-183, 01/10/print 2002.

[109] H. Ge, "UPA, a universal protein array system for quantitative detection of protein–protein, protein–DNA, protein–RNA and protein–ligand interactions," *Nucleic Acids Research,* vol. 28, no. 2, pp. e3-e3, 08/10/received

08/31/revised

10/21/accepted 2000.

[110] J. D. Keene, J. M. Komisarow, and M. B. Friedersdorf, "RIP-Chip: the isolation and identification of mRNAs, microRNAs and protein components of ribonucleoprotein complexes from cell extracts," *Nat. Protocols,* 10.1038/nprot.2006.47 vol. 1, no. 1, pp. 302-307, 06//print 2006.

[111] P. Uetz *et al.*, "A comprehensive analysis of protein-protein interactions in Saccharomyces cerevisiae," *Nature,* 10.1038/35001009 vol. 403, no. 6770, pp. 623-627, 02/10/print 2000.

[112] J. Zahiri, J. H. Bozorgmehr, and A. Masoudi-Nejad, "Computational Prediction of Protein–Protein Interaction Networks: Algorithms and Resources," *Current Genomics,* vol. 14, no. 6, pp. 397-414, 09/

06/12/received

08/07/revised

08/26/accepted 2013.

[113] A. Pan, C. Lahiri, A. Rajendiran, and B. Shanmugham, "Computational analysis of protein interaction networks for infectious diseases," *Briefings in Bioinformatics,* August 10, 2015 2015.

[114] H. Jeong, S. P. Mason, A. L. Barabasi, and Z. N. Oltvai, "Lethality and centrality in protein networks," *Nature,* 10.1038/35075138 vol. 411, no. 6833, pp. 41-42, 05/03/print 2001.

[115] L. H. Hartwell, J. J. Hopfield, S. Leibler, and A. W. Murray, "From molecular to modular cell biology," *Nature,* 1999.

[116] A. J. Berenstein, J. Piñero, L. I. Furlong, and A. Chernomoretz, "Mining the Modular Structure of Protein Interaction Networks," *PLoS ONE,* vol. 10, no. 4, p. e0122477, 2015.

[117] M. J. Cowley *et al.*, "PINA v2.0: mining interactome modules," *Nucleic Acids Research,* vol. 40, no. D1, pp. D862-D865, January 1, 2012 2012.

[118] B.-J. Breitkreutz, C. Stark, and M. Tyers, "Osprey: a network visualization system," *Genome Biol,* vol. 4, no. 3, p. R22, 2003.

[119] R. Saito *et al.*, "A travel guide to Cytoscape plugins," *Nature methods,* vol. 9, no. 11, pp. 1069-1076, 11/06 2012.

[120] K. Han, B. Park, H. Kim, J. Hong, and J. Park, "HPID: The Human Protein Interaction Database," *Bioinformatics,* vol. 20, no. 15, pp. 2466-2470, October 12, 2004 2004.

[121] J. Y. Chen, S. Mamidipalli, and T. Huan, "HAPPI: an online database of comprehensive human annotated and predicted protein interactions," *BMC genomics,* vol. 10, no. Suppl 1, p. S16, 2009.

[122] D. Pratt *et al.*, "NDEx, the Network Data Exchange," *Cell Systems,* vol. 1, no. 4, pp. 302-305, 2015.

[123] S. Mathivanan *et al.*, "An evaluation of human protein-protein interaction data in the public domain," *BMC Bioinformatics,* vol. 7, no. Suppl 5, p. S19, 2006.

[124] A. Platzer, P. Perco, A. Lukas, and B. Mayer, "Characterization of protein-interaction networks in tumors," *BMC Bioinformatics,* journal article vol. 8, no. 1, p. 224, June 27 2007.

[125] J. Sun and Z. Zhao, "A comparative study of cancer proteins in the human protein-protein interaction network," *BMC Genomics,* journal article vol. 11, no. 3, p. S5, December 01 2010.

[126] Y. Yang, L. Han, Y. Yuan, J. Li, N. Hei, and H. Liang, "Gene co-expression network analysis reveals common system-level properties of prognostic genes across cancer types," Article vol. 5, p. 3231, 02/03/online 2014.

[127] D. V. Brown *et al.*, "Coexpression analysis of CD133 and CD44 identifies Proneural and Mesenchymal subtypes of glioblastoma multiforme," *Oncotarget,* vol. 6, no. 8, pp. 6267-6280, 01/31

12/06/received

01/12/accepted 2015.

[128] S. M. Pulst, "Genetic linkage analysis," *Archives of Neurology,* vol. 56, no. 6, pp. 667-672, 1999.

[129] W. S. Bush and J. H. Moore, "Chapter 11: Genome-Wide Association Studies," *PLoS Computational Biology,* vol. 8, no. 12, p. e1002822, 12/27 2012.

[130] D. Guala and E. L. L. Sonnhammer, "A large-scale benchmark of gene prioritization methods," *Scientific Reports,* vol. 7, p. 46598, 04/21

06/09/received

03/22/accepted 2017.

[131] M. Oti, S. Ballouz, and M. A. Wouters, "Web tools for the prioritization of candidate disease genes," (in eng), *Methods Mol Biol,* vol. 760, pp. 189-206, 2011.

[132] Y. Bromberg, "Chapter 15: Disease Gene Prioritization," *PLOS Computational Biology,* vol. 9, no. 4, p. e1002902, 2013.

[133] C. Perez-Iratxeta, P. Bork, and M. A. Andrade, "Association of genes to genetically inherited diseases using data mining," *Nat Genet,* 10.1038/ng895 vol. 31, no. 3, pp. 316-319, 07//print 2002.

[134] P. Zhang, J. Zhang, H. Sheng, J. J. Russo, B. Osborne, and K. Buetow, "Gene functional similarity search tool (GFSST)," *BMC Bioinformatics,* journal article vol. 7, no. 1, p. 135, March 14 2006.

[135] S. Köhler, S. Bauer, D. Horn, and P. N. Robinson, "Walking the Interactome for Prioritization of Candidate Disease Genes," *The American Journal of Human Genetics,* vol. 82, no. 4, pp. 949-958, 2008/04/11/ 2008.

[136] J. Chen, B. J. Aronow, and A. G. Jegga, "Disease candidate gene identification and prioritization using protein interaction networks," *BMC Bioinformatics,* journal article vol. 10, no. 1, p. 73, February 27 2009.

[137] X. Wang, N. Gulbahce, and H. Yu, "Network-based methods for human disease gene prediction," *Briefings in Functional Genomics,* vol. 10, no. 5, pp. 280-293, 2011.

[138] S. Navlakha and C. Kingsford, "The power of protein interaction networks for associating genes with diseases," (in eng), *Bioinformatics,* vol. 26, no. 8, pp. 1057-63, Apr 15 2010.

[139] Y. Yoshida *et al.*, "PosMed (Positional Medline): prioritizing genes with an artificial neural network comprising medical documents to accelerate positional cloning," *Nucleic Acids Research,* vol. 37, no. suppl_2, pp. W147-W152, 2009.

[140] Y. Li and J. C. Patra, "Genome-wide inferring gene-phenotype relationship by walking on the heterogeneous network," *Bioinformatics,* vol. 26, no. 9, pp. 1219-24, May 01 2010.

[141] X. Zeng, Y. Liao, Y. Liu, and Q. Zou, "Prediction and Validation of Disease Genes Using HeteSim Scores," (in eng), *IEEE/ACM Trans Comput Biol Bioinform,* vol. 14, no. 3, pp. 687-695, May-Jun 2017.

[142] C. J. Wolfe, I. S. Kohane, and A. J. Butte, "Systematic survey reveals general applicability of "guilt-by-association" within gene coexpression networks," *BMC Bioinformatics,* journal article vol. 6, no. 1, p. 227, September 14 2005.

[143] S. Oliver, "Proteomics: Guilt-by-association goes global," *Nature,* 10.1038/35001165 vol. 403, no. 6770, pp. 601-603, 02/10/print 2000.

[144] D. Altshuler, M. Daly, and L. Kruglyak, "Guilt by association," *Nat Genet,* 10.1038/79839 vol. 26, no. 2, pp. 135-137, 10//print 2000.

[145] M. Oti and H. G. Brunner, "The modular nature of genetic diseases," *Clinical Genetics,* vol. 71, no. 1, pp. 1-11, 2007.

[146] B. Chen, J. Shi, S. Zhang, and F.-X. Wu, "Identifying protein complexes in protein–protein interaction networks by using clique seeds and graph entropy," *PROTEOMICS,* vol. 13, no. 2, pp. 269-277, 2013.

[147] M. Ayati, S. Erten, M. R. Chance, and M. Koyutürk, "MOBAS: identification of disease-associated protein subnetworks using modularity-based scoring," *EURASIP Journal on Bioinformatics and Systems Biology,* journal article vol. 2015, no. 1, p. 7, June 30 2015.

[148] X. Wu, R. Jiang, M. Q. Zhang, and S. Li, "Network-based global inference of human disease genes," *Molecular systems biology,* vol. 4, no. 1, p. 189, 2008.

[149] X. Zhou, J. Menche, A.-L. Barabási, and A. Sharma, "Human symptoms–disease network," Article vol. 5, p. 4212, 06/26/online 2014.

[150] S. Brohee and J. van Helden, "Evaluation of clustering algorithms for protein-protein interaction networks," *BMC bioinformatics,* vol. 7, no. 1, p. 488, 2006.

[151] R. Sharan, I. Ulitsky, and R. Shamir, "Network-based prediction of protein function," *Molecular Systems Biology,* vol. 3, no. 1, 2007.

[152] B. Adamcsek, G. Palla, I. J. Farkas, I. Derényi, and T. Vicsek, "CFinder: locating cliques and overlapping modules in biological networks," *Bioinformatics,* vol. 22, no. 8, pp. 1021-1023, April 15, 2006 2006.

[153] M. Altaf-Ul-Amin, Y. Shinbo, K. Mihara, K. Kurokawa, and S. Kanaya, "Development and implementation of an algorithm for detection of protein complexes in large interaction networks," *BMC Bioinformatics,* vol. 7, pp. 207-207, 04/14

11/12/received

04/14/accepted 2006.

[154] X. Li, M. Wu, C. K. Kwoh, and S. K. Ng, "Computational approaches for detecting protein complexes from protein interaction networks: a survey," *BMC Genomics,* vol. 11, 2010.

[155] G. D. Bader and C. W. V. Hogue, "An automated method for finding molecular complexes in large protein interaction networks," *BMC Bioinformatics,* vol. 4, 2003.

[156] S. Tripathi, S. Moutari, M. Dehmer, and F. Emmert-Streib, "Comparison of module detection algorithms in protein networks and investigation of the biological meaning of predicted modules," *BMC Bioinformatics,* journal article vol. 17, no. 1, p. 129, March 18 2016.

[157] M. E. Levine, P. Langfelder, and S. Horvath, "A Weighted SNP Correlation Network Method for Estimating Polygenic Risk Scores," in *Biological Networks and Pathway Analysis*, T. V. Tatarinova and Y. Nikolsky, Eds. New York, NY: Springer New York, 2017, pp. 277-290.

[158] F. Hormozdiari, O. Penn, E. Borenstein, and E. E. Eichler, "The discovery of integrated gene networks for autism and related disorders," *Genome Research,* vol. 25, no. 1, pp. 142-154, January 1, 2015 2015.

[159] P. Jia, S. Zheng, J. Long, W. Zheng, and Z. Zhao, "dmGWAS: dense module searching for genome-wide association studies in protein-protein interaction networks," (in eng), *Bioinformatics,* vol. 27, no. 1, pp. 95-102, Jan 01 2011.

[160] C. Ma, Z. Zhao, T. Gui, Y. Chen, X. Dang, and D. Wilkins, "A generative Bayesian model to identify cancer driver genes," in *Bioinformatics and Biomedicine (BIBM), 2015 IEEE International Conference on*, 2015, pp. 351-356: IEEE.

[161] M. Re and G. Valentini, "Cancer module genes ranking using kernelized score functions," *BMC Bioinformatics,* vol. 13, no. Suppl 14, pp. S3-S3, 09/07 2012.

[162] J. Menche *et al.*, "Uncovering disease-disease relationships through the incomplete interactome," *Science,* vol. 347, no. 6224, 2015.

[163] J. Song and M. Singh, "How and when should interactome-derived clusters be used to predict functional modules and protein function?," *Bioinformatics,* vol. 25, no. 23, pp. 3143-3150, 2009.

[164] L. Cowen, T. Ideker, B. J. Raphael, and R. Sharan, "Network propagation: a universal amplifier of genetic associations," (in eng), *Nat Rev Genet,* vol. 18, no. 9, pp. 551-562, Sep 2017.

[165] O. Vanunu and R. Sharan, "A Propagation-based Algorithm for Inferring Gene-Disease Assocations," in *German Conference on Bioinformatics*, 2008, pp. 54-52.

[166] O. Vanunu, O. Magger, E. Ruppin, T. Shlomi, and R. Sharan, "Associating genes and protein complexes with disease via network propagation," *PLoS computational biology,* vol. 6, no. 1, p. e1000641, 2010.

[167] L. Page, S. Brin, R. Motwani, and T. Winograd, "The PageRank citation ranking: Bringing order to the web," Stanford InfoLab1999.

[168] J. SHRAGER, T. HOGG, and B. A. HUBERMAN, "Observation of Phase Transitions in Spreading Activation Networks," *Science,* vol. 236, no. 4805, pp. 1092-1094, 1987.

[169] J. M. Kleinberg, "Authoritative sources in a hyperlinked environment," *Journal of the ACM (JACM),* vol. 46, no. 5, pp. 604-632, 1999.

[170] S. Fortunato, "Community detection in graphs," *Physics reports,* vol. 486, no. 3, pp. 75-174, 2010.

[171] D. J. Klein and M. Randić, "Resistance distance," *Journal of mathematical chemistry,* vol. 12, no. 1, pp. 81-95, 1993.

[172] D.-Y. Cho, Y.-A. Kim, and T. M. Przytycka, "Chapter 5: Network Biology Approach to Complex Diseases," *PLOS Computational Biology,* vol. 8, no. 12, p. e1002820, 2012.

[173] K. Mitra, A.-R. Carvunis, S. K. Ramesh, and T. Ideker, "Integrative approaches for finding modular structure in biological networks," *Nat Rev Genet,* Review vol. 14, no. 10, pp. 719-732, 10//print 2013.

[174] W. S. Noble, R. Kuang, C. Leslie, and J. Weston, "Identifying remote protein homologs by network propagation," *FEBS Journal,* vol. 272, no. 20, pp. 5119-5128, 2005.

[175] P. Csermely, T. Korcsmáros, H. J. M. Kiss, G. London, and R. Nussinov, "Structure and dynamics of molecular networks: A novel paradigm of drug discovery: A comprehensive review," *Pharmacology & Therapeutics,* vol. 138, no. 3, pp. 333-408, 2013/06/01/ 2013.

[176] H. Kitano, "Cancer as a robust system: implications for anticancer therapy," *Nat Rev Cancer,* 10.1038/nrc1300 vol. 4, no. 3, pp. 227-235, 03//print 2004.

[177] A. S. Azmi, B. Bao, and F. H. Sarkar, "Exosomes in Cancer Development, Metastasis and Drug Resistance: A Comprehensive Review," *Cancer metastasis reviews,* vol. 32, no. 0, pp. 10.1007/s10555-013-9441-9, 2013.

[178] T. Tian, S. Olson, J. M. Whitacre, and A. Harding, "The origins of cancer robustness and evolvability," (in eng), *Integr Biol (Camb),* vol. 3, no. 1, pp. 17-30, Jan 2011.

[179] B. Alberts, A. Johnson, J. Lewis, M. Raff, K. Roberts, and P. Walter, "Carrier proteins and active membrane transport," 2002.

[180] E. Mahon, A. Salvati, F. Baldelli Bombelli, I. Lynch, and K. A. Dawson, "Designing the nanoparticle–biomolecule interface for “targeting and therapeutic delivery”," *Journal of Controlled Release,* vol. 161, no. 2, pp. 164-174, 2012/07/20/ 2012.

[181] J. D. Ramsey and N. H. Flynn, "Cell-penetrating peptides transport therapeutics into cells," *Pharmacology & Therapeutics,* vol. 154, pp. 78-86, 10// 2015.

[182] S. E. Andaloussi, I. Mäger, X. O. Breakefield, and M. J. A. Wood, "Extracellular vesicles: biology and emerging therapeutic opportunities," (in eng), *Nat Rev Drug Discov,* vol. 12, no. 5, pp. 347-57, May 2013.

[183] H. Kalra, G. Drummen, and S. Mathivanan, "Focus on Extracellular Vesicles: Introducing the Next Small Big Thing," *International Journal of Molecular Sciences,* vol. 17, no. 2, p. 170, 2016.

[184] L. Balaj *et al.*, "Tumour microvesicles contain retrotransposon elements and amplified oncogene sequences," Article vol. 2, p. 180, 02/01/online 2011.

[185] B.-T. Pan and R. M. Johnstone, "Fate of the transferrin receptor during maturation of sheep reticulocytes in vitro: selective externalization of the receptor," (in eng), *Cell,* vol. 33, no. 3, pp. 967-78, Jul 1983.

[186] H. Valadi, K. Ekstrom, A. Bossios, M. Sjostrand, J. J. Lee, and J. O. Lotvall, "Exosome-mediated transfer of mRNAs and microRNAs is a novel mechanism of genetic exchange between cells," (in eng), *Nat Cell Biol,* vol. 9, no. 6, pp. 654-9, Jun 2007.

[187] M. Record, K. Carayon, M. Poirot, and S. Silvente-Poirot, "Exosomes as new vesicular lipid transporters involved in cell–cell communication and various

pathophysiologies," *Biochimica et Biophysica Acta (BBA) - Molecular and Cell Biology of Lipids,* vol. 1841, no. 1, pp. 108-120, 1// 2014.

[188] B. Février, D. Vilette, H. Laude, and G. Raposo, "Exosomes: a bubble ride for prions?," *Traffic,* vol. 6, no. 1, pp. 10-17, 2005.

[189] K. C. Vallabhaneni *et al.*, "Extracellular vesicles from bone marrow mesenchymal stem/stromal cells transport tumor regulatory microRNA, proteins, and metabolites," *Oncotarget,* vol. 6, no. 7, pp. 4953-4967, 12/31

09/15/received

12/27/accepted 2015.

[190] A. Cvjetkovic *et al.*, "Detailed Analysis of Protein Topology of Extracellular Vesicles–Evidence of Unconventional Membrane Protein Orientation," Article vol. 6, p. 36338, 11/08/online 2016.

[191] S. Mathivanan, H. Ji, and R. J. Simpson, "Exosomes: Extracellular organelles important in intercellular communication," *Journal of Proteomics,* vol. 73, no. 10, pp. 1907-1920, 2010/09/10/ 2010.

[192] E. Cocucci and J. Meldolesi, "Ectosomes and exosomes: shedding the confusion between extracellular vesicles," *Trends in Cell Biology,* vol. 25, no. 6, pp. 364-372, 2015.

[193] S. Elmore, "Apoptosis: A Review of Programmed Cell Death," *Toxicologic pathology,* vol. 35, no. 4, pp. 495-516, 2007.

[194] R. J. Simpson, J. W. E. Lim, R. L. Moritz, and S. Mathivanan, "Exosomes: proteomic insights and diagnostic potential," *Expert Review of Proteomics,* vol. 6, no. 3, pp. 267-283, 2009/06/01 2009.

[195] J. G. van den Boorn, J. Daßler, C. Coch, M. Schlee, and G. Hartmann, "Exosomes as nucleic acid nanocarriers," *Advanced Drug Delivery Reviews,* vol. 65, no. 3, pp. 331-335, 2013/03/01/ 2013.

[196] B. S. Batista, W. S. Eng, K. T. Pilobello, K. D. Hendricks-Muñoz, and L. K. Mahal, "Identification of a Conserved Glycan Signature for Microvesicles," *Journal of Proteome Research,* vol. 10, no. 10, pp. 4624-4633, 2011/10/07 2011.

[197] T. B. Steinbichler, J. Dudás, H. Riechelmann, and I.-I. Skvortsova, "The role of exosomes in cancer metastasis," *Seminars in Cancer Biology,* vol. 44, no. Supplement C, pp. 170-181, 2017/06/01/ 2017.

[198] J. L. Hood, R. S. San, and S. A. Wickline, "Exosomes Released by Melanoma Cells Prepare Sentinel Lymph Nodes for Tumor Metastasis," *Cancer Research,* vol. 71, no. 11, pp. 3792-3801, 2011.

[199] C. Grange *et al.*, "Microvesicles released from human renal cancer stem cells stimulate angiogenesis and formation of lung premetastatic niche," (in eng), *Cancer Res,* vol. 71, no. 15, pp. 5346-56, Aug 01 2011.

[200] D. W. Greening *et al.*, "Emerging roles of exosomes during epithelial–mesenchymal transition and cancer progression," *Seminars in Cell & Developmental Biology,* vol. 40, no. Supplement C, pp. 60-71, 2015/04/01/ 2015.

[201] D. Ha, N. Yang, and V. Nadithe, "Exosomes as therapeutic drug carriers and delivery vehicles across biological membranes: current perspectives and future challenges," *Acta Pharmaceutica Sinica B,* vol. 6, no. 4, pp. 287-296, 2016/07/01/ 2016.

[202] A. Aryani and B. Denecke, "Exosomes as a Nanodelivery System: a Key to the Future of Neuromedicine?," *Molecular Neurobiology,* journal article vol. 53, no. 2, pp. 818-834, March 01 2016.

[203] F. Properzi, M. Logozzi, and S. Fais, "Exosomes: the future of biomarkers in medicine," *Biomarkers,* vol. 7, no. 5, pp. 769-778, 2013.

[204] J. Kowal, M. Tkach, and C. Théry, "Biogenesis and secretion of exosomes," *Current Opinion in Cell Biology,* vol. 29, pp. 116-125, 2014/08/01/ 2014.

[205] M. Colombo *et al.*, "Analysis of ESCRT functions in exosome biogenesis, composition and secretion highlights the heterogeneity of extracellular vesicles," *Journal of Cell Science,* vol. 126, no. 24, pp. 5553-5565, 2013.

[206] O. Schmidt and D. Teis, "The ESCRT machinery," *Current Biology,* vol. 22, no. 4, pp. R116-R120, 2012.

[207] S. Stuffers, C. Sem Wegner, H. Stenmark, and A. Brech, "Multivesicular Endosome Biogenesis in the Absence of ESCRTs," *Traffic,* vol. 10, no. 7, pp. 925-937, 2009.

[208] D. Perez-Hernandez *et al.*, "The intracellular interactome of tetraspanin-enriched microdomains reveals their function as sorting machineries toward exosomes," (in eng), *J Biol Chem,* vol. 288, no. 17, pp. 11649-61, Apr 26 2013.

[209] A. H. Hutagalung and P. J. Novick, "Role of Rab GTPases in Membrane Traffic and Cell Physiology," *Physiological reviews,* vol. 91, no. 1, pp. 119-149, 2011.

[210] M. Babst, "MVB Vesicle Formation: ESCRT-Dependent, ESCRT-Independent and Everything in Between," *Current opinion in cell biology,* vol. 23, no. 4, pp. 452-457, 05/11 2011.

[211] J. Votteler and Wesley I. Sundquist, "Virus Budding and the ESCRT Pathway," *Cell Host & Microbe,* vol. 14, no. 3, pp. 232-241, 2013/09/11/ 2013.

[212] A. Jemal, F. Bray, M. M. Center, J. Ferlay, E. Ward, and D. Forman, "Global cancer statistics," *CA: A Cancer Journal for Clinicians,* vol. 61, no. 2, pp. 69-90, 2011.

[213] T. C. G. Atlas, "Comprehensive molecular characterization of human colon and rectal cancer," *Nature,* 10.1038/nature11252 vol. 487, no. 7407, pp. 330-337, 07/19/print 2012.

[214] Cancer Genome Atlas Network, "Comprehensive molecular characterization of human colon and rectal cancer," *Nature,* vol. 487, no. 7407, pp. 330-337, 2012.

[215] A. Sadanandam *et al.*, "A colorectal cancer classification system that associates cellular phenotype and responses to therapy," *Nat Med,* vol. 19, no. 5, pp. 619-25, May 2013.

[216] B. Zhang *et al.*, "Proteogenomic characterization of human colon and rectal cancer," *Nature,* vol. 513, no. 7518, pp. 382-7, Sep 18 2014.

[217] S. A. Forbes *et al.*, "COSMIC: exploring the world's knowledge of somatic mutations in human cancer," *Nucleic Acids Res,* vol. 43, no. Database issue, pp. D805-11, Jan 2015.

[218] M. S. Lawrence *et al.*, "Discovery and saturation analysis of cancer genes across 21 tumour types," *Nature,* Article vol. 505, no. 7484, pp. 495-501, 01/23/print 2014.

[219] G. Gundem *et al.*, "IntOGen: integration and data mining of multidimensional oncogenomic data," *Nat Meth,* 10.1038/nmeth0210-92 vol. 7, no. 2, pp. 92-93, 02//print 2010.

[220] O. An, V. Pendino, M. D'Antonio, E. Ratti, M. Gentilini, and F. D. Ciccarelli, "NCG 4.0: the network of cancer genes in the era of massive mutational screenings of cancer genomes," *Database,* vol. 2014, January 1, 2014 2014.

[221] M. Zhao, J. Sun, and Z. Zhao, "TSGene: a web resource for tumor suppressor genes," *Nucleic Acids Research,* vol. 41, no. D1, pp. D970-D976, January 1, 2013 2013.

[222] Z. Shi, J. Wang, and B. Zhang, "NetGestalt: integrating multidimensional omics data over biological networks," *Nat Methods,* vol. 10, no. 7, pp. 597-8, Jul 2013.

[223] J. Zhu, Z. Shi, J. Wang, and B. Zhang, "Empowering biologists with multi-omics data: colorectal cancer as a paradigm," *Bioinformatics,* vol. 31, no. 9, pp. 1436-43, May 1 2015.

[224] J. Barretina *et al.*, "The Cancer Cell Line Encyclopedia enables predictive modelling of anticancer drug sensitivity," *Nature,* 10.1038/nature11003 vol. 483, no. 7391, pp. 603-307, 03/29/print 2012.

[225] M. Uhlen *et al.*, "Proteomics. Tissue-based map of the human proteome," *Science,* vol. 347, no. 6220, p. 1260419, Jan 23 2015.

[226] S. Mathivanan *et al.*, "Human Proteinpedia enables sharing of human protein data," *Nat Biotechnol,* vol. 26, no. 2, pp. 164-7, Feb 2008.

[227] T. S. Keshava Prasad *et al.*, "Human Protein Reference Database—2009 update," *Nucleic Acids Research,* vol. 37, no. Database issue, pp. D767-D772, 11/06

09/16/received

10/20/revised

10/22/accepted 2009.

[228] M. C. Chambers *et al.*, "A cross-platform toolkit for mass spectrometry and proteomics," *Nat Biotech,* Opinion and Comment vol. 30, no. 10, pp. 918-920, 10//print 2012.

[229] R. Craig and R. C. Beavis, "TANDEM: matching proteins with tandem mass spectra," *Bioinformatics,* vol. 20, no. 9, pp. 1466-1467, June 12, 2004 2004.

[230] A. C. Paoletti *et al.*, "Quantitative proteomic analysis of distinct mammalian Mediator complexes using normalized spectral abundance factors," *Proc Natl Acad Sci U S A,* vol. 103, no. 50, pp. 18928-33, Dec 12 2006.

[231] S. Mathivanan, H. Ji, B. J. Tauro, Y. S. Chen, and R. J. Simpson, "Identifying mutated proteins secreted by colon cancer cell lines using mass spectrometry," *J Proteomics,* vol. 76 Spec No., pp. 141-9, Dec 5 2012.

[232] S. Keerthikumar *et al.*, "Proteogenomic analysis reveals exosomes are more oncogenic than ectosomes," *Oncotarget,* vol. 6, no. 17, pp. 15375-96, Jun 20 2015.

[233] S. Keerthikumar *et al.*, "ExoCarta: A web-based compendium of exosomal cargo," *J Mol Bio,* 2015.

[234] Z. Wang, B. Vogelstein, and K. W. Kinzler, "Phosphorylation of beta-catenin at S33, S37, or T41 can occur in the absence of phosphorylation at T45 in colon cancer cells," *Cancer Res,* vol. 63, no. 17, pp. 5234-5, Sep 1 2003.

[235] M. Fasolini, X. Wu, M. Flocco, J. Y. Trosset, U. Oppermann, and S. Knapp, "Hot spots in Tcf4 for the interaction with beta-catenin," *J Biol Chem,* vol. 278, no. 23, pp. 21092-8, Jun 6 2003.

[236] A. Chatr-aryamontri *et al.*, "The BioGRID interaction database: 2015 update," *Nucleic Acids Research,* vol. 43, no. D1, pp. D470-D478, January 28, 2015 2015.

[237] M. Milacic *et al.*, "Annotating Cancer Variants and Anti-Cancer Therapeutics in Reactome," *Cancers,* vol. 4, no. 4, p. 1180, 2012.

[238] M. Kanehisa, S. Goto, Y. Sato, M. Kawashima, M. Furumichi, and M. Tanabe, "Data, information, knowledge and principle: back to metabolism in KEGG," *Nucleic Acids Research,* vol. 42, no. D1, pp. D199-D205, January 1, 2014 2014.

[239] J. M. Mariadason *et al.*, "Gene expression profiling-based prediction of response of colon carcinoma cells to 5-fluorouracil and camptothecin," *Cancer Res,* vol. 63, no. 24, pp. 8791-812, Dec 15 2003.

[240] M. Pathan *et al.*, "FunRich: An open access standalone functional enrichment and interaction network analysis tool," *Proteomics,* vol. 15, no. 15, pp. 2597-601, Aug 2015.

[241] F. Emmert-Streib *et al.*, "Functional and genetic analysis of the colon cancer network," *BMC Bioinformatics,* journal article vol. 15, no. 6, p. S6, May 16 2014.

[242] M. Pathan *et al.*, "A novel community driven software for functional enrichment analysis of extracellular vesicles data," *Journal of Extracellular Vesicles,* vol. 6, no. 1, p. 1321455, 2017/01/01 2017.

[243] S. Segditsas and I. Tomlinson, "Colorectal cancer and genetic alterations in the Wnt pathway," *Oncogene,* vol. 25, no. 57, pp. 7531-7537, //print 2006.

[244] D. J. Watts and S. H. S, "Collective dynamics of 'small-world' networks," *Nature,* vol. 393, 1998.

[245] R. Sever and J. S. Brugge, "Signal Transduction in Cancer," *Cold Spring Harbor Perspectives in Medicine,* vol. 5, no. 4, April 1, 2015 2015.

[246] T. R. Medler and L. M. Coussens, "Duality of the Immune Response in Cancer: Lessons Learned from Skin," *The Journal of investigative dermatology,* vol. 134, no. E1, pp. E23-E28, 10/10 2014.

[247] E.-h. Jho, T. Zhang, C. Domon, C.-K. Joo, J.-N. Freund, and F. Costantini, "Wnt/β-catenin/Tcf signaling induces the transcription of Axin2, a negative regulator of the signaling pathway," *Molecular and cellular biology,* vol. 22, no. 4, pp. 1172-1183, 2002.

[248] Z.-Q. Wu *et al.*, "Canonical Wnt suppressor, Axin2, promotes colon carcinoma oncogenic activity," *Proceedings of the National Academy of Sciences,* vol. 109, no. 28, pp. 11312-11317, July 10, 2012 2012.

[249] M. Zitt *et al.*, "Dickkopf-3 as a new potential marker for neoangiogenesis in colorectal cancer: expression in cancer tissue and adjacent non-cancerous tissue," *Disease markers,* vol. 24, no. 2, pp. 101-109, 2008.

[250] K. Birkenkamp-Demtröder *et al.*, "Keratin23 (KRT23) Knockdown Decreases Proliferation and Affects the DNA Damage Response of Colon Cancer Cells," *PLOS ONE,* vol. 8, no. 9, p. e73593, 2013.

[251] L. Lin *et al.*, "STAT3 Is Necessary for Proliferation and Survival in Colon Cancer–Initiating Cells," *Cancer Research,* vol. 71, no. 23, pp. 7226-7237, 2011.

[252] K.-U. Wagner *et al.*, "Tsg101 Is Essential for Cell Growth, Proliferation, and Cell Survival of Embryonic and Adult Tissues," *Molecular and Cellular Biology,* vol. 23, no. 1, pp. 150-162, January 1, 2003 2003.

[253] H. Lan *et al.*, "APOBEC3G expression is correlated with poor prognosis in colon carcinoma patients with hepatic metastasis," *International Journal of Clinical and Experimental Medicine,* vol. 7, no. 3, pp. 665-672, 03/15

01/17/received

02/20/accepted 2014.

[254] H. L. Huang *et al.*, "Silencing of argininosuccinate lyase inhibits colorectal cancer formation," (in eng), *Oncol Rep,* vol. 37, no. 1, pp. 163-170, Jan 2017.

[255] T. Abbas and A. Dutta, "p21 in cancer: intricate networks and multiple activities," *Nature reviews. Cancer,* vol. 9, no. 6, pp. 400-414, 2009.

[256] A. Palaniappan, K. Ramar, and S. Ramalingam, "Computational Identification of Novel Stage-Specific Biomarkers in Colorectal Cancer Progression," *PLOS ONE,* vol. 11, no. 5, p. e0156665, 2016.

[257] H. Y. Peng *et al.*, "Knockdown of ELMO3 Suppresses Growth, Invasion and Metastasis of Colorectal Cancer," (in eng), *Int J Mol Sci,* vol. 17, no. 12, Dec 16 2016.

[258] E. Ziolko *et al.*, "The profile of melatonin receptors gene expression and genes associated with their activity in colorectal cancer: a preliminary report," (in eng), *J Biol Regul Homeost Agents,* vol. 29, no. 4, pp. 823-8, Oct-Dec 2015.

[259] J. Gao *et al.*, "Association of NFKBIA polymorphism with colorectal cancer risk and prognosis in Swedish and Chinese populations," (in eng), *Scand J Gastroenterol,* vol. 42, no. 3, pp. 345-50, Mar 2007.

[260] L. E. King, C. G. Love, O. M. Sieber, M. C. Faux, and A. W. Burgess, "Differential RNA-seq analysis comparing APC-defective and APC-restored SW480 colorectal cancer cells," *Genomics Data,* vol. 7, pp. 293-296, 3// 2016.

[261] M. D. Robinson, D. J. McCarthy, and G. K. Smyth, "edgeR: a Bioconductor package for differential expression analysis of digital gene expression data," *Bioinformatics,* vol. 26, no. 1, pp. 139-140, 2010.

[262] Y. Chen, A. T. Lun, and G. K. Smyth, "Differential expression analysis of complex RNA-seq experiments using edgeR," in *Statistical analysis of next generation sequencing data*: Springer, 2014, pp. 51-74.

[263] S. A. Forbes *et al.*, "COSMIC: somatic cancer genetics at high-resolution," *Nucleic Acids Research,* vol. 45, no. D1, pp. D777-D783, 2017.

[264] T. Hart *et al.*, "High-Resolution CRISPR Screens Reveal Fitness Genes and Genotype-Specific Cancer Liabilities," *Cell,* vol. 163, no. 6, pp. 1515-1526, 2015/12/03/ 2015.

[265] T. Wang *et al.*, "Identification and characterization of essential genes in the human genome," *Science,* vol. 350, no. 6264, pp. 1096-1101, 2015.

[266] T. Bertomeu *et al.*, "A High-Resolution Genome-Wide CRISPR/Cas9 Viability Screen Reveals Structural Features and Contextual Diversity of the Human Cell-Essential Proteome," *Molecular and Cellular Biology,* vol. 38, no. 1, January 1, 2018 2018.

[267] G. Alanis-Lobato, M. A. Andrade-Navarro, and M. H. Schaefer, "HIPPIE v2.0: enhancing meaningfulness and reliability of protein–protein interaction networks," *Nucleic Acids Research,* vol. 45, no. D1, pp. D408-D414, 2017.

[268] P. Futreal *et al.*, "A census of human cancer genes," *Nat Rev Cancer,* vol. 4, 2004.

[269] J. Eric, O. Travis, P. Pearu, and others, "SciPy: Open Source Scientific Tools for Python," ed, 2001-.

[270] D. Chisanga, S. Keerthikumar, S. Mathivanan, and N. Chilamkurti, "Integration of heterogeneous ‘omics’ data using semi-supervised network labelling to identify essential genes in colorectal cancer," *Computers & Electrical Engineering,* vol. 67, pp. 267-277, 4// 2018.

[271] R. P. Horgan and L. C. Kenny, "'Omic' technologies: genomics, transcriptomics, proteomics and metabolomics," *The Obstetrician & Gynaecologist,* vol. 13, no. 3, pp. 189-195, 2011.

[272] Z. Dong and Y. Chen, "Transcriptomics: advances and approaches," (in eng), *Sci China Life Sci,* vol. 56, no. 10, pp. 960-7, Oct 2013.

[273] A. Stencel and B. Crespi, "What is a genome?," *Molecular Ecology,* vol. 22, no. 13, pp. 3437-3443, 2013.

[274] L. B. Holder, M. M. Haque, and M. K. Skinner, "Machine learning for epigenetics and future medical applications," (in eng), *Epigenetics,* pp. 1-10, May 19 2017.

[275] S. Keerthikumar, "An Introduction to Proteome Bioinformatics," in *Proteome Bioinformatics*, S. Keerthikumar and S. Mathivanan, Eds. New York, NY: Springer New York, 2017, pp. 1-3.

[276] D. S. Wishart, R. Mandal, A. Stanislaus, and M. Ramirez-Gaona, "Cancer Metabolomics and the Human Metabolome Database," (in eng), *Metabolites,* vol. 6, no. 1, Mar 02 2016.

[277] J. T. Erler and R. Linding, "Network-based drugs and biomarkers," (in eng), *J Pathol,* vol. 220, no. 2, pp. 290-6, Jan 2010.

[278] J. Chen *et al.*, "Protein interaction network underpins concordant prognosis among heterogeneous breast cancer signatures," *J. of Biomedical Informatics,* vol. 43, no. 3, pp. 385-396, 2010.

[279] M. Hofree, J. P. Shen, H. Carter, A. Gross, and T. Ideker, "Network-based stratification of tumor mutations," *Nat Meth,* Article vol. 10, no. 11, pp. 1108-1115, 11//print 2013.

[280] X. Zhu, M. Gerstein, and M. Snyder, "Getting connected: analysis and principles of biological networks," *Genes & Development,* vol. 21, no. 9, pp. 1010-1024, May 1, 2007 2007.

[281] L. Ou-Yang, D.-Q. Dai, X.-L. Li, M. Wu, X.-F. Zhang, and P. Yang, "Detecting temporal protein complexes from dynamic protein-protein interaction networks," *BMC bioinformatics,* vol. 15, no. 1, p. 335, 2014.

[282] R. Jansen *et al.*, "A Bayesian networks approach for predicting protein-protein interactions from genomic data," *Science,* vol. 302, no. 5644, pp. 449-453, 2003.

[283] M. A. Pujana *et al.*, "Network modeling links breast cancer susceptibility and centrosome dysfunction," *Nature genetics,* vol. 39, no. 11, pp. 1338-1349, 2007.

[284] H. J. Cordell, "Detecting gene-gene interactions that underlie human diseases," (in eng), *Nat Rev Genet,* vol. 10, no. 6, pp. 392-404, Jun 2009.

[285] T. Gui, X. Dong, R. Li, Y. Li, and Z. Wang, "Identification of hepatocellular carcinoma-related genes with a machine learning and network analysis," (in eng), *J Comput Biol,* vol. 22, no. 1, pp. 63-71, Jan 2015.

[286] F. Crick, "Central dogma of molecular biology," (in eng), *Nature,* vol. 227, no. 5258, pp. 561-3, Aug 8 1970.

[287] A. Hagberg, P. Swart, and D. S Chult, "Exploring network structure, dynamics, and function using NetworkX," Los Alamos National Laboratory (LANL)2008.

[288] D. Zhou, O. Bousquet, T. N. Lal, J. Weston, and B. Schölkopf, "Learning with local and global consistency," in *Advances in neural information processing systems*, 2004, pp. 321-328.

[289] M. Ruffalo, M. Koyutürk, and R. Sharan, "Network-based integration of disparate omic data to identify" silent players" in cancer," *PLoS computational biology,* vol. 11, no. 12, p. e1004595, 2015.

[290] S. E. Craig and S. M. Brady-Kalnay, "Cancer cells cut homophilic cell adhesion molecules and run," (in eng), *Cancer Res,* vol. 71, no. 2, pp. 303-9, Jan 15 2011.

[291] U. Cavallaro and G. Christofori, "Cell adhesion and signalling by cadherins and Ig-CAMs in cancer," *Nature reviews. Cancer,* vol. 4, no. 2, p. 118, 2004.

[292] T. Bose, A. Cieslar-Pobuda, and E. Wiechec, "Role of ion channels in regulating Ca(2)(+) homeostasis during the interplay between immune and cancer cells," (in eng), *Cell Death Dis,* vol. 6, p. e1648, Feb 19 2015.

[293] G. R. Monteith, D. McAndrew, H. M. Faddy, and S. J. Roberts-Thomson, "Calcium and cancer: targeting Ca2+ transport," *Nature reviews. Cancer,* vol. 7, no. 7, p. 519, 2007.

[294] J. I. Fletcher, M. Haber, M. J. Henderson, and M. D. Norris, "ABC transporters in cancer: more than just drug efflux pumps," (in eng), *Nat Rev Cancer,* vol. 10, no. 2, pp. 147-56, Feb 2010.

[295] F. Leonessa and R. Clarke, "ATP binding cassette transporters and drug resistance in breast cancer," (in eng), *Endocr Relat Cancer,* vol. 10, no. 1, pp. 43-73, Mar 2003.

[296] T. Agarwal, N. Annamalai, T. K. Maiti, and H. Arsad, "Biophysical changes of ATP binding pocket may explain loss of kinase activity in mutant DAPK3 in cancer: A molecular dynamic simulation analysis," (in eng), *Gene,* vol. 580, no. 1, pp. 17-25, Apr 10 2016.

[297] S. I. Yun *et al.*, "Ubiquitin specific protease 4 positively regulates the WNT/beta-catenin signaling in colorectal cancer," (in eng), *Mol Oncol,* vol. 9, no. 9, pp. 1834-51, Nov 2015.

[298] Y. Wang *et al.*, "Ubiquitin-specific protease 14 (USP14) regulates cellular proliferation and apoptosis in epithelial ovarian cancer," (in eng), *Med Oncol,* vol. 32, no. 1, p. 379, Jan 2015.

[299] M. I. Davis *et al.*, "Small Molecule Inhibition of the Ubiquitin-specific Protease USP2 Accelerates cyclin D1 Degradation and Leads to Cell Cycle Arrest in Colorectal Cancer and Mantle Cell Lymphoma Models," (in eng), *J Biol Chem,* vol. 291, no. 47, pp. 24628-24640, Nov 18 2016.

[300] A. Pal, M. A. Young, and N. J. Donato, "Emerging potential of therapeutic targeting of ubiquitin-specific proteases in the treatment of cancer," (in eng), *Cancer Res,* vol. 74, no. 18, pp. 4955-66, Sep 15 2014.

[301] S. P. Tsofack *et al.*, "NONO and RALY proteins are required for YB-1 oxaliplatin induced resistance in colon adenocarcinoma cell lines," (in eng), *Mol Cancer,* vol. 10, p. 145, Nov 25 2011.

[302] A. H. Sillars-Hardebol *et al.*, "TPX2 and AURKA promote 20q amplicon-driven colorectal adenoma to carcinoma progression," (in eng), *Gut,* vol. 61, no. 11, pp. 1568-75, Nov 2012.

[303] S. A. Best, A. N. Nwaobasi, C. D. Schmults, and M. R. Ramsey, "CCAR2 Is Required for Proliferation and Tumor Maintenance in Human Squamous Cell Carcinoma," (in eng), *J Invest Dermatol,* vol. 137, no. 2, pp. 506-512, Feb 2017.

[304] P. D. Robbins and A. E. Morelli, "Regulation of immune responses by extracellular vesicles," *Nat Rev Immunol,* Review vol. 14, no. 3, pp. 195-208, 03//print 2014.

[305] L. Barile and G. Vassalli, "Exosomes: Therapy delivery tools and biomarkers of diseases," *Pharmacology & Therapeutics,* vol. 174, pp. 63-78, 6// 2017.

[306] S. Mathivanan, J. W. E. Lim, B. J. Tauro, H. Ji, R. L. Moritz, and R. J. Simpson, "Proteomics Analysis of A33 Immunoaffinity-purified Exosomes Released from the Human Colon Tumor Cell Line LIM1215 Reveals a Tissue-specific Protein Signature," *Molecular & Cellular Proteomics,* vol. 9, no. 2, pp. 197-208, February 1, 2010 2010.

[307] C. Raiborg and H. Stenmark, "The ESCRT machinery in endosomal sorting of ubiquitylated membrane proteins," *Nature,* 10.1038/nature07961 vol. 458, no. 7237, pp. 445-452, 03/26/print 2009.

[308] N. Deo, *Graph theory with applications to engineering and computer science*. Courier Dover Publications, 2016.

[309] J. Scott, "Social network analysis: developments, advances, and prospects," *Social network analysis and mining,* vol. 1, no. 1, pp. 21-26, 2011.

[310] D. Arrell and A. Terzic, "Network systems biology for drug discovery," *Clinical Pharmacology & Therapeutics,* vol. 88, no. 1, pp. 120-125, 2010.

[311] F. Cheng *et al.*, "Prediction of drug-target interactions and drug repositioning via network-based inference," *PLoS Comput Biol,* vol. 8, no. 5, p. e1002503, 2012.

[312] K. M. Borgwardt, C. S. Ong, S. Schönauer, S. Vishwanathan, A. J. Smola, and H.-P. Kriegel, "Protein function prediction via graph kernels," *Bioinformatics,* vol. 21, no. suppl 1, pp. i47-i56, 2005.

[313] J. B. Pereira-Leal, A. J. Enright, and C. A. Ouzounis, "Detection of functional modules from protein interaction networks," *PROTEINS: Structure, Function, and Bioinformatics,* vol. 54, no. 1, pp. 49-57, 2004.

[314] O. Keskin, N. Tuncbag, and A. Gursoy, "Predicting protein–protein interactions from the molecular to the proteome level," *Chemical reviews,* vol. 116, no. 8, pp. 4884-4909, 2016.

[315] A. Vinayagam *et al.*, "Integrating protein-protein interaction networks with phenotypes reveals signs of interactions," *Nature methods,* vol. 11, no. 1, pp. 94-99, 2014.

[316] J. Gillis, S. Ballouz, and P. Pavlidis, "Bias tradeoffs in the creation and analysis of protein–protein interaction networks," *Journal of Proteomics,* vol. 100, pp. 44-54, 4/4/ 2014.

[317] M. A. W. Oortveld *et al.*, "Human Intellectual Disability Genes Form Conserved Functional Modules in Drosophila," *PLOS Genetics,* vol. 9, no. 10, p. e1003911, 2013.

[318] E. L. L. Sonnhammer and G. Östlund, "InParanoid 8: orthology analysis between 273 proteomes, mostly eukaryotic," *Nucleic Acids Research,* vol. 43, no. D1, pp. D234-D239, 2015.

[319] A. Chatr-aryamontri *et al.*, "The BioGRID interaction database: 2017 update," *Nucleic Acids Research,* vol. 45, no. D1, pp. D369-D379, 2017.

[320] T. S. Keshava Prasad *et al.*, "Human Protein Reference Database—2009 update," *Nucleic Acids Research,* vol. 37, no. suppl 1, pp. D767-D772, January 1, 2009 2009.

[321] M. Ashburner *et al.*, "Gene Ontology: tool for the unification of biology," *Nature genetics,* vol. 25, no. 1, pp. 25-29, 2000.

[322] J. Z. Wang, Z. Du, R. Payattakool, P. S. Yu, and C.-F. Chen, "A new method to measure the semantic similarity of GO terms," *Bioinformatics,* vol. 23, no. 10, pp. 1274-1281, 2007.

[323] G. Yu, F. Li, Y. Qin, X. Bo, Y. Wu, and S. Wang, "GOSemSim: an R package for measuring semantic similarity among GO terms and gene products," (in eng), *Bioinformatics,* vol. 26, no. 7, pp. 976-8, Apr 01 2010.

[324] M. F. Baietti *et al.*, "Syndecan–syntenin–ALIX regulates the biogenesis of exosomes," *Nature Cell Biology,* Article vol. 14, p. 677, 06/03/online 2012.

[325] R. Agarwal, B. Kumar, M. Jayadev, D. Raghav, and A. Singh, "CoReCG: a comprehensive database of genes associated with colon-rectal cancer," *Database: The Journal of Biological Databases and Curation,* vol. 2016, p. baw059, 04/25

06/24/received

02/28/revised

03/23/accepted 2016.

[326] M. A. Mahdavi and Y.-H. Lin, "False positive reduction in protein-protein interaction predictions using gene ontology annotations," *BMC Bioinformatics,* vol. 8, pp. 262-262, 07/23

03/21/received

07/23/accepted 2007.

Appendix

Supplementary table 4.1

List of predicted genes with their associated local area connectivity, differential gene expression status and z-score.

Supplementary table 6.1

List of predicted ESCRT neighbours that changed the physical coherence of the ESCRT network. The list consists of proteins that either increased or decreased the physical coherence of the network and enriched for cytoplasm and/or cytosol.

Chapter 14

Network Tools for the Analysis of Proteomic Data

David Chisanga, Shivakumar Keerthikumar, Suresh Mathivanan,
and Naveen Chilamkurti

Abstract

Recent advancements in high-throughput technologies such as mass spectrometry have led to an increase in the rate at which data is generated and accumulated. As a result, standard statistical methods no longer suffice as a way of analyzing such gigantic amounts of data. Network analysis, the evaluation of how nodes relate to one another, has over the years become an integral tool for analyzing high throughput proteomic data as they provide a structure that helps reduce the complexity of the underlying data.

Computational tools, including pathway databases and network building tools, have therefore been developed to store, analyze, interpret, and learn from proteomics data. These tools enable the visualization of proteins as networks of signaling, regulatory, and biochemical interactions. In this chapter, we provide an overview of networks and network theory fundamentals for the analysis of proteomics data. We further provide an overview of interaction databases and network tools which are frequently used for analyzing proteomics data.

Key words Proteomics, Network theory, Protein–protein interactions, Network tools, Network analysis, Bioinformatics

1 Introduction

In recent years, the development of high-throughput technologies such as next-generation sequencing techniques in the field of genomics and tandem mass spectrometry in the field of proteomics and metabolomics has led to the birth of the “omics” study [1]. These techniques and tools involved in the study of functional genomics and other omics data have constantly helped in our understanding of cellular biology and have drastically reduced the cost of conducting “omics” related studies. The speed with which data are generated and disseminated today means that researchers can gain insight for the fraction of the cost compared to that in past years. For instance, by using tandem mass spectrometry, two groups [2, 3] have developed the first draft of the human proteome. Also, using bioinformatics, another group integrated

Shivakumar Keerthikumar and Suresh Mathivanan (eds.), *Proteome Bioinformatics*, Methods in Molecular Biology, vol. 1549, DOI 10.1007/978-1-4939-6740-7_14, © Springer Science+Business Media LLC 2017

publicly available proteomics datasets to map 96% of the human proteome [1].

However, with terabytes of proteomic data pouring into research centers every day, standard statistical methods for analyzing data are becoming ineffective. Researchers are faced with the formidable task of how to take advantage of this heterogeneous data to gain insight in areas such as disease and drug development as well as answering questions such as the following: How can they characterize and manipulate complex interactome of basic elements such as genes and proteins? How can they visualize these interactomes and infer meaningful information from them?

Network theory has long played a fundamental role in disciplines ranging from computer science, sociology, engineering, and physics, to molecular and population biology [4]. In biology and medicine, network analysis methods are applied in areas such as drug target identification, prediction of a gene or protein function, protein complex or module detection, prediction of novel interactions and functional associations, identification of disease subnetworks, disease biomarker identification, and mapping of disease pathways [5]. Networks have long been used in a variety of fields to reduce the complexity of data [6, 7]. Computational tools, including pathway databases and network building tools, have been developed to store, analyze, and interpret biological networks [8].

This chapter provides an overview of the application of network theory in analyzing and visualization of proteomic data by discussing various tools used for storage, analysis, and interpretation of proteomic data through the use of biological networks with an emphasis on protein–protein interaction networks. To get started, we provide a brief background to both proteomics and network theory.

1.1 Background to Proteomics

Coined by Marc Wilkins and colleagues [9] in the mid-1990s to mimic the terms "genomics" and "genome," respectively, proteomics is in essence a systems science whose aim is to identify and record the functions as well as structures of proteins in organisms. Proteomics is a systems science which involves not only the measurement of proteins but also the measurement of their expressions in a cell and the interplay of proteins, protein complexes, signaling pathways, and network modules.

Proteins are termed as the workhorses of cellular systems, as they perform an array of cellular functions ranging from catalyzing reactions, cellular transportation, transcription of DNA information to RNA, and acting as molecular motors to signaling [10]. They perform these functions not on their own, but within large complexes where they interact with other molecules like proteins, DNA, RNA as well as with other small molecules. Because of their importance, a malfunction in key proteins can lead to serious pathological outcomes like cancer, metabolic imbalances, and neurodegenerative diseases. With significant ongoing research into protein functionality and their interactions with other molecules in understanding disease, research has turned to network theory concepts to model and study these interactions.

1.2 Background to Network Theory Concepts

A network or a graph (in mathematics) is a collection of objects connected by lines. The objects are called nodes or vertices while the connections between the objects are called edges or links and are drawn as lines between points as shown in Fig. 1

Formally, a network is a graph G defined as an ordered pair $G=(V, E)$ where V is a set of nodes and E is a set of edges [4]. Nodes are said to be adjacent if they are joined by an edge while node 'A' is said to be a neighbor to node 'B' if adjacent to node 'B' and vice versa. Edges between nodes can be undirected (Fig. 1) or directed (Fig. 2), as such a graph $G=(V, E)$ is called undirected if

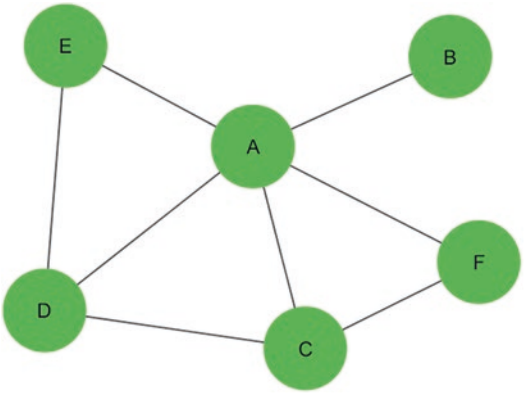


Fig. 1 Shows an example of an undirected network graph in which each node is connected by an edge that does not show the origin and destination by way of an ***arrow***

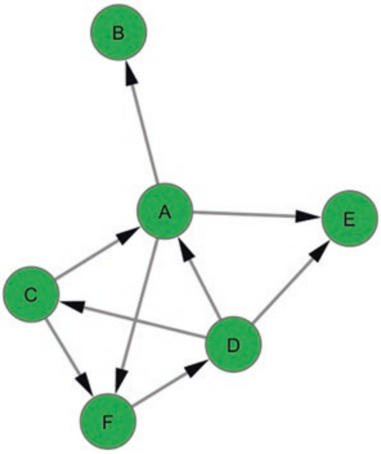


Fig. 2 Shows an example of a directed graph in which each node is connected by an edge with an ***arrow*** indicative of the direction of the relationship

an edge vv' (where v and v' are nodes in set *V*) in set E of edges implies that it is the same as edge v'v also in *E*; otherwise *G* is called directed. A directed acyclic graph, on the other hand, is a directed graph that contains no cycles. Finally, a graph is said to be connected if there is a path from any node to any other node.

Using the above network/graph concepts, researchers have used networks to reduce the complexity of systems thereby making it easier to draw conclusions from them. Networks are applied in various fields such as computer networks, social networks, and interactome networks in molecular biology research.

Interactome networks provide a global picture that is useful in understanding how interactions between molecules influence cellular behavior [11]. It has been established that biological behavior arises from the complex interactions between the cell's numerous molecules such as proteins, DNA, RNA, and other small molecules. Common examples of interactomes in molecular biology are; protein–protein interactions, virus–host networks, transcriptional regulatory networks, metabolic networks, and disease networks. Protein–protein interactions (PPIs) form the backbone of signaling pathways, metabolic pathways, and cellular processes required for normal functioning of cells [12].

The steps to perform proteomic analysis can be summed up by use of a flowchart as shown in Fig. 3, it involves identifying a set of target proteins of biological interest needs to be studied and then followed by retrieval or identification of interacting partners from

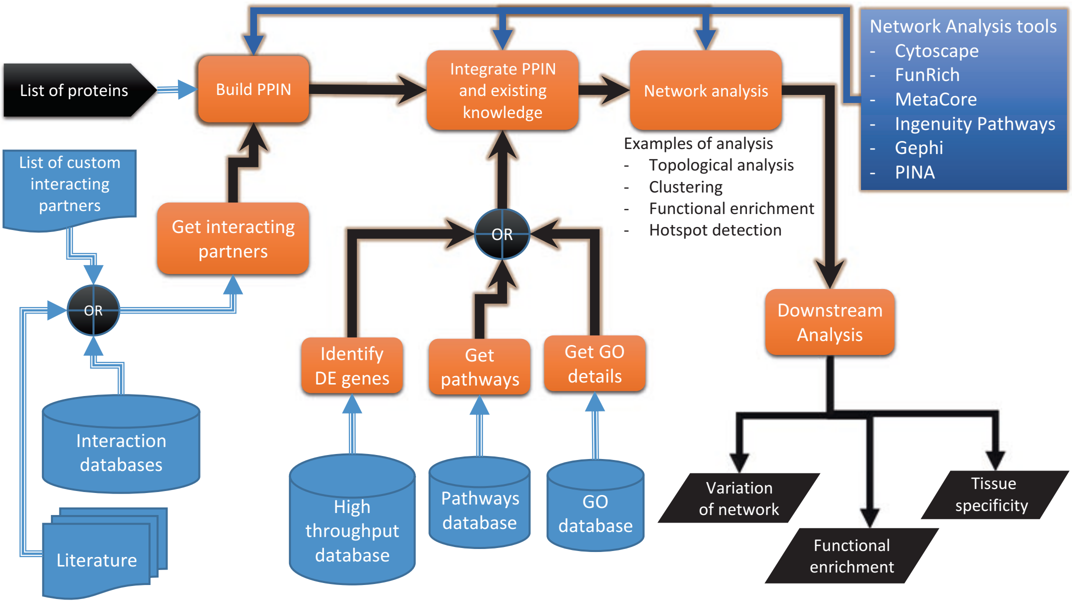


Fig. 3 Shows a summary representation of the steps involved in analyzing proteomic data using network theory concepts. It also shows the data types required and from where they can be sourced. It also gives an example of expected outputs from the network analysis

various interaction resources discussed below. An interaction network is then generated and integrated with any existing knowledge such as gene ontology (GO) enrichment, biological pathways or differential gene or protein expression. A topological analysis of the network is then performed using metrics such as degree, degree centrality or betweenness centrality which is further followed on by downstream analysis to identify network variations, functional enrichment of identified modules, or tissue specificity.

2 Protein–Protein Interaction Databases

The mappings of proteins and their interacting partners have been curated by various groups and deposited into online databases. These databases are typically Web-based resources that serve as archives of information pertaining to the mapping of protein interactions, functional enrichment (GO enrichment) and pathway details. These databases act as sources of protein mapping information in network analysis. The most widely used PPI databases include Human Protein Reference Database (HPRD) [13], Molecular Interaction Database (MINT) [14], Biological General Repository for Interaction Database (BioGRID) [15], Search Tool for Recurring Instances of Neighboring Genes/Proteins (STRING) [16], Database of Interacting Proteins (DIP) [17], Biomolecular Interaction Network Database (BIND) [18], and the IntAct molecular interaction database (IntAct). Depending on the database, the annotations may be based on experimental observations while other databases such as STRING can have a high proportion of predicted and literature mined interactions. Below, we briefly discuss the most commonly used databases while Table 1 provides a summary of these database resources with protein–protein interaction mappings.

2.1 BioGRID

The Biological General Repository for Interaction Datasets (BioGRID) is an open, accessible Web-based repository of genetic and protein interaction mappings which have been curated from the primary biomedical literature of humans and other major model organism species [15]. As of May 2016, the database housed over 1,000,000 protein and genetic interactions curated from over 56,000 high-throughput datasets and individually focused publications for major model organisms.

BioGRID features an easy to use Web interface with a search tool which users can use to search against the database, the search results then show the interacting partners, interactor details, and a graphical network visualization of the interacting partners. Users can then manipulate the network by either changing the network layout or filtering through the network by node degrees. In addition, users can also download custom defined or entire interaction

Table 1
Summary of database resources that house protein–protein interactions and their respective features

Resource	Description	URL link	Reference	No. proteins	No. interactions	No. organisms
BIND	Biomolecular Interaction Network Database	http://bond.unleashedinformatics.com/	[18]	23,643	43,050	80
BioGRID	Biological General Repository for Interaction Datasets	http://thebiogrid.org/	[15]	56,105	553,827	175
HPRD	Human Protein Reference Database	http://www.hprd.org	[13]	30,047	41,327	1
IntAct	IntAct Molecular Interaction Database	http://www.ebi.ac.uk/intact/	[20]	89,716	356,806	131
MINT	Molecular INTeraction database	http://mint.bio.uniroma2.it/mint	[14]	35,553	241,458	144
STRING	Search Tool for the Retrieval of Interacting Genes/Proteins	http://string-db.org/	[16]	9,643,763		2031

datasets for offline network analysis and downstream analysis. BioGRID also features online tools and resources that allow for the use of BioGRID data. A number of visualization tools such as Osprey, Cytoscape, and GeneMania, data management tools like ProHits, plugins like BioGRID Tab File Loader Plugin for Cytoscape and BiogridPlugin2 for Cytoscape as well as Web services BioGRID REST Service and PSICQUIC provide users with access to or can be used to analyze BioGRID data.

2.2 Human Protein Reference Database

Human Protein Reference Database is a Web-based resource that houses experimentally derived human proteome information [13]. It is one of the most comprehensive collections of human proteome information resource available online. It houses information pertaining to; protein–protein interactions, posttranslational modifications and tissue expression. As of May, 2016, the database housed over 30,000 protein entries, over 41,000 protein–protein interactions, 93,000 posttranslational modifications (PTMs), 112,000 protein expressions, 22,000 subcellular localization details, 400 domains and with over 453,000 PubMed links to publications.

The landing page of HPRD provides a range of features ranging from a querying functionality, BLAST feature to a browse feature. Users can query the database using the query page through a number of search options, the results are then displayed using graphical visual displays and are categorized into protein information, PTMs, protein length, and protein–protein interactions. Users can similarly get protein information through the browse page where the information is grouped into molecular classes, domains, motifs, PTMs and based on localization. HPRD further includes a Basic Alignment Search Tool (BLAST) which allows users to search against the database based on the provided protein or nucleotide sequence. Other features included are a phosphor motif finder tool which searches across user submitted protein sequence for the presence of over 300 phosphorylation-based motifs listed in HPRD. HPRD also provides tab delimited files for binary protein–protein interactions which users can download for offline processing and further download stream analysis.

2.3 Molecular INTeraction Database (MINT)

The Molecular INTeraction database [19] is a Web-based resource that stores physical interactions between proteins of model organisms that have been curated from the scientific literature. As of May 2016, MINT had over 241,000 protein–protein interactions, 35,000 proteins, and over 5000 PubMed links to publications.

MINT data can be downloaded in several formats such as PSI-ML, tab-delimited and MINT flat file formats. Otherwise, users can use the search feature that allows users to search the MINT database. Users can search the database using several options such as by gene name, protein accession number, or any 6-character keyword. A user defined list of proteins can furthermore be uploaded and used to generate a network visualization based on the information in the database.

2.4 Biomolecular Interaction Network Database

The Biomolecular Interaction Network Database [18] is a Web-based resource for PPI data and was one of the earliest resources for biomolecular interactions (proteins, genes, etc.), molecular complexes and pathways. BIND initiated by the University of Toronto as part of the Biomolecular Object Network Databank (BOND) has since been acquired by Thomson Reuters. BIND provides tools for data specification plus a database which is accompanied by data mining and visualization tools.

2.5 IntAct Molecular Interaction Database

IntAct [20] is an open-source Web-based molecular interaction database that catalogs data curated from the scientific literature or from direct data depositions. As of May 2016, IntAct had over 591,000 molecular interactions, and 91,000 interactors sourced from over 14,000 publications.

Using IntAct users can explore the fine details of the mechanism by which a specific protein binds to protein partners or use the entire interactome of an organism to perform a network analysis of large-scale 'omics' experiment. The front-end of IntAct features a search tool that can be used to search against the IntAct database. Users can then view the interacting partners, interaction details and a graphical presentation of the network.

2.6 Search Tool for Recurring Instances of Neighboring Genes/Proteins (STRING)

STRING is a freely available Web-based biological database that houses information on experimentally derived and predicted protein–protein interactions for a number of organisms. This information has been curated from various sources, including experimental data, computational prediction methods, and published literature. STRING holds over 184 million interactions, 9,000,000 proteins from over 2000 organisms.

STRING provides an easy-to-use Web interface that allows users to quickly search for a protein of interest and visualize and download interaction data. It further has a Cytoscape plugin which allows users to directly access the STRING database from Cytoscape. The interaction data returned from STRING is weighted and allows for the calculation of confidence scores for each interaction. In addition, STRING has capabilities that allow it to connect to other databases and consequently perform literature mining. It also includes a capability that allows for the drawing of simple protein networks based on the provided list of genes and the available interactions in the database.

3 PPI Data Exchange Formats

Interaction networks are represented in a number of different file formats, the most widely used formats are; tab delimited text (.tab or .txt format), excel workbooks (.xls format), simple interaction file (SIF or .sif format), nested network format (NNF or .nff format), graph markup language (GML or .gml format), XGMML (extensible graph markup and modeling language), SBML, BioPAX, PSI-MI level 1 and 2.5 formats. All the interaction repositories provide at least one of these formats as a way to download interaction data.

3.1 Delimited Text and Excel Workbooks

The delimited text and excel workbook file formats are the most basic and widely used for working with interactive data and are supported by most if not all network analysis tools. Tables in these files can contain network and edge (interaction) attributes or values such as the confidence of an interaction. With these types of files, users can specify the columns for source and target nodes as well as interaction types, and edge attributes when importing network data into an analysis tool.

3.2 Simple Interaction Format (SIF)

This format allows for the construction of a network from a list of interactions by easily merging different interaction sets into a larger network.

Each line of an SIF file annotates a source node, a relationship (or edge type), and one or more target nodes as shown in the following example:

```
nodeA <relationship type> nodeB
nodeC <relationship type> nodeB
nodeD <relationship type> nodeA
```

3.3 Nested Network Format

This format is simple and similar to the SIF format except it allows the option of nesting a network into a single a node. An interaction is specified by either of two possible formats [21, 22]:

- A node “node” contained in a “network”:
  - Network node.
- Two nodes linked together contained in a network:
  - Network node1 interaction with node2.

3.4 Graph Markup Language (GML)

GML unlike the SIF format comes with a language that supports rich graph formatting and is widely supported by most visualization software tools. A GML formatted file can contain information pertaining to graphs, nodes, and edges, and hence capable of emulating almost every other format. A network can be built using the SIF format and by applying network layouts can then be stored as a GML file as this preserves the layout of a network. Further details on the GML specification can be found on the GML documentation website: http://www.fim.uni-passau.de/index.php?id=17297&L=1.

Other formats such as XGMML is the XML extension of the GML format and is the preferred format to GML, Systems Biology Markup Language (SMBL) format is an XML format used to describe biochemical networks, the specification for SMBL can be found on the website: http://sbml.org/Documents/Specifications, PSI-ML format specification is an XML-based format that is used for data exchange of protein–protein interactions. GraphML is another XML-based format for generating graphs. Apart from the XML-based formats, JSON-based file formats are increasingly being used for data exchange of protein–protein interactions (Subheading 2.3).

4 Network Analysis and Visualization Tools

This section discusses some of the commonly used tools in the proteomics network analysis, but before delving into what tools to use, we begin this discussion by looking at the ways by which networks can be quantified in order to provide more informative results.

4.1 Quantifying Networks

The most commonly applied metric are; degree, degree distribution, scale-free networks, the degree exponent, shortest path, mean path length, and clustering coefficient [23]. By using these network metrics, we can quantify and characterize important network features which are not commonly visible.

Protein–protein interactions are the most commonly used form of networks in proteomic data analysis. In these networks, proteins are represented as nodes while interactions between the nodes are depicted by edges or links. This mapping of proteins is based on experimental information which has been obtained from methods such as mass spectrometer [24], protein chip technologies [25, 26], yeast two-hybrid screens [27], and predictions from computational methods [28]. These mappings have been collected and deposited into online databases as discussed below.

Network tools are mainly used to analyze proteomic data through functional annotation, knowledge integration, modularity analysis, topological analysis, and basic network property analysis [29].

The basic properties of a network such as node degree, degree distribution, betweenness centrality, and eigenvector centrality can be used to deduce the significance of a protein [30]. Another important metric is the identification of modules which represent a vital level of organization in biology [31]. A module in proteomics can be defined as a set of interacting proteins that can be associated with a common biological process. By using networks, clusters of interacting proteins can be identified as modules and associated with a functionality. Modules provide a comprehensive and global description of interaction patterns to comprehend the complexity of biological systems [32]. Module detection enables functional annotation of constituent proteins and the discovery of targets for therapy in diseases such as cancer. In addition to detection of modules, the integration of existing knowledge into networks plays a vital role in the analysis of proteomic data. Such knowledge may include integrating Gene Ontology (GO) annotations, differential gene expression, and pathway details. By highlighting such information, candidate disease proteins may be identified and module functions can be annotated.

4.2 Steps to Performing Network Analysis

To perform network analysis on proteomic data, there are a number of steps that are involved; these steps are summarized in Fig. 3. The steps involved include but are not limited to:

1. The first step involves identifying a list of proteins or genes that need to be analyzed using a network tool. The researcher can select which protein or gene appears on the lists, as per individual needs.
2. Interacting partners of these proteins are then obtained from any of the databases discussed above.
3. A protein–protein interaction network is then built by using a visualizing tool from the tools listed in Table 2.
4. To get more meaningful information from the network, the protein–protein interaction network is then integrated with already existing knowledge such as pathways, differential expressions for genes or proteins obtained from either high-throughput custom data or online databases such as The Cancer Genome Atlas (TCGA). Other existing knowledge that can be integrated includes Gene Ontology enrichment, which can help to identify the functional annotations of the modules or individual proteins in the network.
5. During topological analysis, network theory concepts such as degree, degree centrality distribution, Eigenvectors, and degree distribution are applied to identify proteins or nodes playing significant roles in the network, variations between a normal and an altered network and modules that can be mapped to a functionality.
6. Topology analysis is further followed by downstream analysis whose objective is mostly dependent on the researcher.
7. Some of the results that may be obtained from a network analysis of proteomic data include a visual representation of the network, module identification, network variations as well as functional enrichment of proteins and modules.

Table 2
Summary of Network tools for analyzing proteomic data

Tool	Reference	URL link	Features
Cytoscape	[22]	http://cytoscape.org/	Open source, Data integration, Network visualization, Network Analysis, Functional enrichment, extensible by plugins, Stand-alone, Platform independent
FunRich (Functional Enrichment Analysis)	[8]	http://funrich.org/	Open source, Functional enrichment, Dataset comparison, Network visualization and analysis, Stand-alone, Runs only on Windows, Results can be exported in various formats
MetaCore	By Thomson Reuters	https://portal.genego.com/	Proprietary, Network visualization, Network analysis, Function enrichment analysis, Data mining toolkit, Network alignment
Ingenuity Pathways Analysis	IPA®, QIAGEN Redwood City	www.qiagen.com/ingenuity	Proprietary, Network visualization and modeling, Causal network analysis, Network analysis, Functional enrichment analysis, Pathway enrichment analysis, Literature mining, Allows for collaboration
Gephi	Gephi	https://gephi.org	Network visualization, Network analysis, Network clustering, Module identification, Dynamic network analysis, Real-time visualization
PINA: Protein Interaction Analysis	[37]	http://cbg.garvan.unsw.edu.au/pina/	Network construction, Module detection, Functional enrichment, Network metric analysis, Network visualization, Community driven annotation
Osprey	[39]	http://biodata.mshri.on.ca/osprey/servlet/Index	Network visualization, Integrates BioGRID, Ability to compare functions between datasets, Build interaction network from custom datasets, Search for specific genes within a network, filtering feature

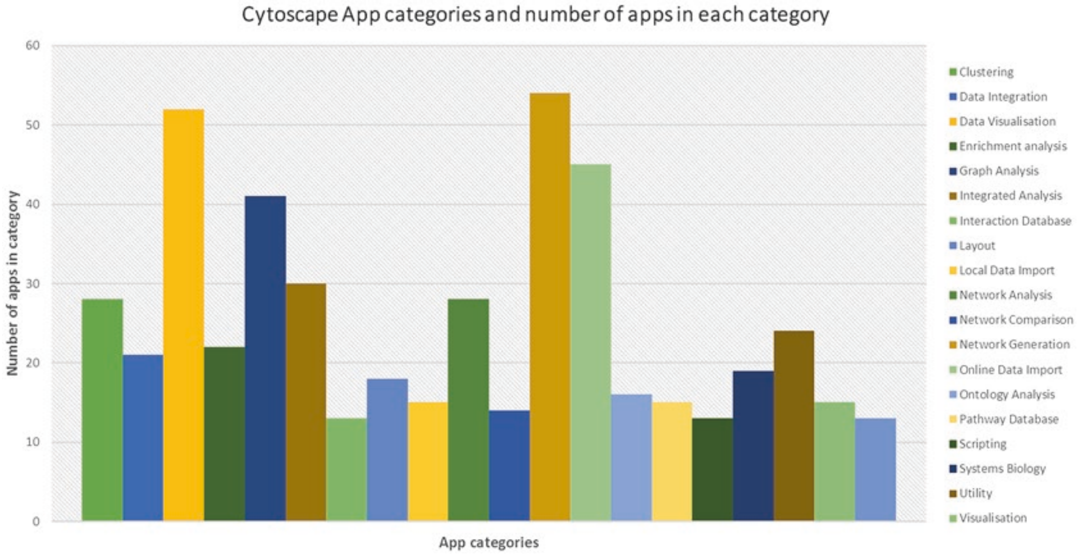


Fig. 4 Shows the distribution of apps or plugins across a number of categories in Cytoscape

4.3 Cytoscape

Cytoscape developed by Trey Ideker (a leading pioneer of systems biology) is a platform independent and open source software tool for the integration, visualization, and statistical modeling of molecular networks together with other systems-level data [21, 33]. The core of Cytoscape provides users with the fundamental features to perform functions such as data integration, analysis, and network visualization. The core also has limited information stored but interconnects with other databases to obtain relevant information. Cytoscape functionality is extensible through the integration of plugins (http://apps.cytoscape.org/) which are now called apps from version 3.0 of Cytoscape.

The apps can be categorized into one or more of the following functional categories such as clustering, data integration, data visualization, enrichment analysis, graph analysis, and integrated analysis. Other functional categories include interaction database, layout, local data import, network analysis, network comparison, network generation, online data import, ontology analysis, pathway database, scripting, systems biology, utility, and visualization. Figure 4 shows the distribution of these apps across the different functional categories.

The first step to a typical Cytoscape workflow is the importation of interactions. These interactions are imported from either a user's own experiment data or from public databases. Data from experiments is loaded directly into Cytoscape using a standard file format such as generic tabular formats including CSV, Excel, and TSV or network-specific formats such as SIF, XGMML, GML, PSI-MI, BioPAX (Biological Pathway Exchange), OpenBEL (Open Biological Expression Language), and SBML.

Importation of data from databases, on the other hand, requires the installation of plugins (apps). A list of genes of interest is passed as a query for interactions from the database. Examples of apps for importing data from databases include the BioGRID database plugin that can be used to import an entire interactome from the BioGRID database. Other ways in which networks can be imported into a network by mining interactions directly from literature or using computational inference from non-interaction data such as expression profiles. This is also achieved through the use of third-party apps. An example of such apps that is Agilent Literature Search software which is a meta-search tool that can automatically search through multiple texts based search engines to extract associations among a set of genes or proteins of interest.

Once the networks are imported into Cytoscape and network visualization is done, network analysis is achieved using the huge collection of apps. For example, using network topology apps like Knowledge-fused Differential Dependency Network (KDDN), users are able to calculate such statistics as network distribution of node degrees. Network clustering apps such as MCODE enable users to extract network regions which are densely connected, thereby forming modules which can then be related to complexes or pathways. Network enrichment apps are used to infer the functions of the identified modules by detecting functional terms that are statistically overrepresented among the set of genes making up the module. Examples of apps that can perform functional enrichment include BiNGO which is a tool that can determine which Gene Ontology categories are statistically overrepresented in a set of genes or a module, the ReactomeFIPlugin is another app that can be used to associate a set of genes in a module to pathways that are related to diseases such as cancer. Furthermore, functional modules can also be identified by integrating networks with expression data to infer network regions that are consistently up- or downregulated. Another example of network analysis that can be done using apps in Cytoscape is network comparison, this involves comparing networks across species or in different conditions to identify regions of the network with conserved interactions. GASOLINE (Greedy and Stochastic algorithm for Optimal Local Alignment of Interaction NEtworks) is an example of an app that can be used to compare multiple networks.

Cytoscape also supports the use of scripting languages such as Python and R. It enables users to develop their own scripts and integrate or call Cytoscape functionality in the order they want it to be done.

4.4 FunRich

Functional Enrichment Analysis (FunRich) tool [8] is an open source stand-alone desktop software tool for functional enrichment and protein–protein interaction network analysis of biological molecules. Features of FunRich include functional enrichment and network analysis of genes and proteins. In addition, FunRich allows the representation of results in editable graphical form as Venn, Bar, Column, Pie and Doughnut charts. FunRich users can perform a biological process, cellular component, molecular function, protein domain, site of expression, biological pathway, transcription, and clinical synopsis phenotypic term enrichment. Users can analyze their datasets against two built-in background databases; FunRich and UniProt or against a customized background database. FunRich does not require users to install any additional applications or plugins to conduct any of the above analysis. FunRich is currently only available for Microsoft's Windows Operating system with plans underway to support other major operating system platforms.

The first step to performing an enrichment analysis in FunRich is the specification of an annotation database. By default, FunRich comes with a human annotation database. Each database consists of biological function annotations and an interaction database. FunRich also comes with the latest UniProt annotation database, otherwise, users can also include a custom database. Once an annotation database has been specified, a list of genes or proteins is then imported. The user can perform a range of analyses on the datasets including comparison across the datasets using a Venn diagram that shows which proteins or genes are common across the datasets. Users can also perform gene set enrichment analysis to determine what biological functions are statically enriched in the gene or protein lists. In addition to these, FunRich also allows users to generate and build an interaction network from where users can then manipulate the network through enriched pathways and functions.

4.5 MetaCore

MetaCore from Thomson Reuters is an integrated proprietary software suite capable of analyzing multiple types of biological data, for example, Next Generation Sequencing [34], variant, Copy Number Variation (CNV), microarray, metabolic, proteomics, microRNA etc. Functional analysis in MetaCore is performed against a high quality, a manually curated database containing molecular interactions vis-à-vis protein–protein interactions, protein–DNA interactions, and protein–RNA interactions. The database is also made up of molecular classes such as transcription factors, signaling and metabolic pathways, and disease ontologies. MetaCore was developed for the purpose of representing biological functionality along with the integration of functional, molecular, or clinical information. Using the data mining toolkit available in MetaCore, users can perform functions like data visualization, analysis, and exchange of data, network alignment using multiple network alignment algorithms, and enrichment analysis. While MetaCore provides a set of rich features, it is a paid for a suite of software for integrated analysis.

4.6 Ingenuity Pathways Analysis

IPA (IPA®, QIAGEN Redwood City, www.qiagen.com/ingenuity) is a proprietary software application with features that allow scientists to model, analyze, and understand the complexity of biological and chemical systems [35]. IPA offers a host of network analysis functions some of these include causal network analysis which allows researchers to identify upstream molecules that control the expression of genes in their datasets and network analysis which allows the building and exploration of transcription of molecular networks such as microRNA, transcriptional networks, and protein–protein interaction networks. Network analysis in IPA can identify regulatory events that lead from signaling events to transcriptional effects, help in understanding toxicity responses by exploring connections between drugs or targets and related genes or chemicals. Users can also edit and expand networks based on the molecular relationships most relevant to the project.

IPA is capable of identifying pathways, molecular mechanisms and biological processes that are relevant to a given dataset. It is also capable of finding biological and chemical knowledge from the scientific literature. Other features allow for collaboration, sharing of results and insights with project teams.

IPA is a subscription-based software application. It is made available as a Web-based, hosted or deployed solution.

4.7 Gephi

Gephi is an open-source data exploratory, network visualization and analysis software tool for large network graphs. Gephi allows users to explore, analyze, spatialize, filter, cluster, manipulate, and export all types of network graphs. With Gephi, users can derive hypotheses and identify patterns by analyzing data using networks.

Gephi can be used to analyze a variety of networks ranging from biological networks to social networks. It supports the majority of the network file formats discussed in Subheading 2.2 above. The core of Gephi can perform basic network metric analysis such as calculating betweenness centrality, closeness, clustering, community detection or module identification. Gephi further includes a feature that allows for the analysis of dynamic networks where a set of networks representing or derived from different conditions or events are compared to infer differences. In addition, Gephi is also extensible by a range of plugins which users can install to perform functionality that is not included in the core of Gephi. While Gephi provides a range of network analysis features, other biological specific network analysis features such as functional enrichment cannot be easily done due to the unavailability of such functionality within Gephi or its associated plugins.

4.8 NDEx-The Network Data Exchange

NDEx-The Network Data Exchange is not so much a network analysis tool, but rather an open source framework for sharing of networks of many types and formats, publication of networks as data, and the use of networks in modular software [36]. Unlike other similar tools such as KEGG and IntAct, NDEx is a data commons framework that allows users to manage the sharing and the publication of networks. Users can upload any type of networks such as pathway models, interaction maps, and novel data-driven knowledge networks. NDEx supports networks of varying formats including simple interaction format (SIF), extensible graph markup and modeling language (XGMML), BioPAX3, and OpenBEL. Each network uploaded to NDEx is given an accession number which acts as a universally unique identifier allowing users to share or include such networks in publications. NDEx also promotes the development of network analysis algorithms and applications by providing access to networks which can be used as inputs through a Web-based relational state transfer application programming interface (REST API). In addition, users can anonymously access networks by searching through the Web interface (www.ndexbio.org). The framework can also be downloaded and run on a local server or personal computer, depending on the needs of a user.

4.9 PINA: Protein Interaction Analysis

Protein Interaction Analysis is a Web-based integrated network analysis platform for protein interaction network construction, filtering, analysis, visualization, and management [37]. PINA has a quarterly updated backend database consisting of an integration of data from six other publicly available databases; IntAct, MINT, BioGRID, DIP, HPRD, and MIPS MPact. To construct a network, PINA provides a query feature where users can either query a single protein, a list of proteins, a list of protein pairs or two lists of proteins.

The constructed PPI networks can be further analyzed by PINA's inbuilt GO term and protein domain annotation tools. Other analyses that can be performed include the use of graph theoretical tools to either discover basic topology properties of a PPI network or identify topologically important proteins, such as hubs or bottlenecks, based on several centrality measures from protein domains and GO terms. In addition, the constructed networks can be downloaded in customized tab-delimited, GraphML or MITAB formats for further analysis using tools such as Cytoscape where they can be integrated with gene expression profiles.

4.10 Colorectal Cancer Atlas

Colorectal Cancer Atlas [38] is an integrated Web-based resource mainly meant for those involved in colorectal cancer research. The tool provides a platform that catalogs both non-quantitative and quantitative proteomic and genomic sequence variation data in both colorectal cancer cell lines and tissues. This information has been curated from existing literature.

Colorectal Cancer Atlas features an easy to use search functionality that also offers auto-complete. Users can search for a given protein, gene, pathway, or cell line that may be of interest to them. Depending the type of search term, the tool then performs functional, pathway, and GO enrichment, maps sequence variances known in colorectal cancer and associated with the searched term, and generates a protein–protein interaction network.

The network integrates proteomic data with genomic sequence variations. Users can use this network analysis module to quickly get an overall picture of the interacting partners of a given gene in colorectal cancer. It uses color intensities to indicate the number of sequence variances for a given gene in the database. Users can also filter through the network by either a gene symbol or by cell lines.

While this tool is specific to colorectal cancer, it provides features that users can quickly use to get functional enrichment information for a given protein or gene as well as perform a gene or protein centered network analysis. Overall, researchers can quickly look up a list of genes or proteins and get an overview of a given gene in colorectal cancer.

4.11 Osprey

Osprey [39] is a software tool that allows for the visualization and analysis of complex interaction networks. Just like most visualization tools, in osprey genes are represented as nodes and interactions as edges. Developed using Java, Osprey is platform independent running on both Linux and Windows based systems.

Osprey provides a range of features that allows users to easily build data-rich graphical representations of their datasets. In addition, users can use the default BioGRID's Gene Ontology interaction datasets to quickly build an interaction network. Some of the features in Osprey include ability to compare functions between datasets, use of custom datasets to build interaction networks, ability to search for specific genes within a network as well filter functions to filter for specific nodes within a large a network. Osprey also has a number of network layouts including concentric circles, spoke, circular, and dual ring, these layouts allow for the comparison of large-scale datasets in an additive manner.

5 Conclusions

In order to study and understand complex systems such as cellular systems, we show that network theory provides metrics that can be used to study such systems using a bottom-up approach. In this chapter, we give an overview of how network theory can be applied to the analysis and study of proteomics data based on a number of network theory metrics. Such metrics include node degree, node centrality, Eigen vector values, and modularity.

We also discuss the most frequently used network analysis tools in analyzing proteomic data. In doing so, a generic workflow that one can use during the analysis is described. Tools discussed include databases which are used to house protein–protein interaction network annotations and the analytical tools that can be applied in analyzing proteomic data.

References

1. Mathivanan S (2014) Integrated bioinformatics analysis of the publicly available protein data shows evidence for 96% of the human proteome. J Proteomics Bioinformatics 2014(7):041–049. doi:10.4172/jpb.1000301
2. Kim M-S, Pinto SM, Getnet D, Nirujogi RS, Manda SS, Chaerkady R, Madugundu AK, Kelkar DS, Isserlin R, Jain S, Thomas JK, Muthusamy B, Leal-Rojas P, Kumar P, Sahasrabuddhe NA, Balakrishnan L, Advani J, George B, Renuse S, Selvan LDN, Patil AH, Nanjappa V, Radhakrishnan A, Prasad S, Subbannayya T, Raju R, Kumar M, Sreenivasamurthy SK, Marimuthu A, Sathe GJ, Chavan S, Datta KK, Subbannayya Y, Sahu A, Yelamanchi SD, Jayaram S, Rajagopalan P, Sharma J, Murthy KR, Syed N, Goel R, Khan AA, Ahmad S, Dey G, Mudgal K, Chatterjee A, Huang T-C, Zhong J, Wu X, Shaw PG, Freed D, Zahari MS, Mukherjee KK, Shankar S, Mahadevan A, Lam H, Mitchell CJ, Shankar SK, Satishchandra P, Schroeder JT, Sirdeshmukh R, Maitra A, Leach SD, Drake CG, Halushka MK, Prasad TSK, Hruban RH, Kerr CL, Bader GD, Iacobuzio-Donahue CA, Gowda H, Pandey A (2014) A draft map of the human proteome. Nature 509(7502):575–581. doi:10.1038/nature13302
3. Wilhelm M, Schlegl J, Hahne H, Gholami AM, Lieberenz M, Savitski MM, Ziegler E, Butzmann L, Gessulat S, Marx H, Mathieson T, Lemeer S, Schnatbaum K, Reimer U, Wenschuh H, Mollenhauer M, Slotta-Huspenina J, Boese J-H, Bantscheff M, Gerstmair A, Faerber F, Kuster B (2014) Mass-spectrometry-based draft of the human proteome. Nature 509(7502):582–587. doi:10.1038/nature13319
4. Pavlopoulos GA, Secrier M, Moschopoulos CN, Soldatos TG, Kossida S, Aerts J, Schneider R, Bagos PG (2011) Using graph theory to analyze biological networks. BioData Mining 4(1):1–27. doi:10.1186/1756-0381-4-10
5. Sevimoglu T, Arga KY (2014) The role of protein interaction networks in systems biomedicine. Comput Struct Biotechnol J 11(18):22–27. doi:10.1016/j.csbj.2014.08.008
6. Gandhi TKB, Zhong J, Mathivanan S, Karthick L, Chandrika KN, Mohan SS, Sharma S, Pinkert S, Nagaraju S, Periaswamy B, Mishra G, Nandakumar K, Shen B, Deshpande N, Nayak R, Sarker M, Boeke JD, Parmigiani G, Schultz J, Bader JS, Pandey A (2006) Analysis of the human protein interactome and comparison with yeast, worm and fly interaction datasets. Nat Genet 38(3):285–293, http://www.nature.com/ng/journal/v38/n3/suppinfo/ng1747_S1.html
7. Mathivanan S, Periaswamy B, Gandhi T, Kandasamy K, Suresh S, Mohmood R, Ramachandra Y, Pandey A (2006) An evaluation of human protein-protein interaction data in the public domain. BMC Bioinformatics 7(5):1–14. doi:10.1186/1471-2105-7-s5-s19
8. Pathan M, Keerthikumar S, Ang C-S, Gangoda L, Quek CYJ, Williamson NA, Mouradov D, Sieber OM, Simpson RJ, Salim A, Bacic A, Hill AF, Stroud DA, Ryan MT, Agbinya JI, Mariadason JM, Burgess AW, Mathivanan S (2015) FunRich: an open access standalone functional enrichment and interaction network analysis tool. Proteomics 15(15):2597–2601. doi:10.1002/pmic.201400515
9. Wilkins MR, Pasquali C, Appel RD, Ou K, Golaz O, Sanchez J-C, Yan JX, Gooley AA, Hughes G, Humphery-Smith I, Williams KL, Hochstrasser DF (1996) From proteins to proteomes: large scale protein identification by two-dimensional electrophoresis and amino acid analysis. Nat Biotechnol 14(1):61–65
10. Schmidt A, Forne I, Imhof A (2014) Bioinformatic analysis of proteomics data. BMC Syst Biol 8(Suppl 2):S3. doi:10.1186/1752-0509-8-S2-S3
11. Blais A, Dynlacht BD (2005) Constructing transcriptional regulatory networks. Genes Dev 19(13):1499–1511
12. De Las RJ, Fontanillo C (2010) Protein–protein interactions essentials: key concepts to building and analyzing interactome networks. PLoS Comput Biol 6(6), e1000807. doi:10.1371/journal.pcbi.1000807
13. Keshava Prasad TS, Goel R, Kandasamy K, Keerthikumar S, Kumar S, Mathivanan S, Telikicherla D, Raju R, Shafreen B, Venugopal A, Balakrishnan L, Marimuthu A, Banerjee S, Somanathan DS, Sebastian A, Rani S, Ray S, Harrys Kishore CJ, Kanth S, Ahmed M, Kashyap MK, Mohmood R, Ramachandra YL, Krishna V, Rahiman BA, Mohan S, Ranganathan P, Ramabadran S, Chaerkady R, Pandey A (2009) Human protein reference database—2009 update. Nucleic Acids Res 37(Database issue):D767–D772. doi:10.1093/nar/gkn892
14. Licata L, Briganti L, Peluso D, Perfetto L, Iannuccelli M, Galeota E, Sacco F, Palma A, Nardozza AP, Santonico E, Castagnoli L, Cesareni G (2012) MINT, the molecular interaction database: 2012 update. Nucleic Acids Res 40(D1):D857–D861. doi:10.1093/nar/gkr930
15. Chatr-Aryamontri A, Breitkreutz BJ, Oughtred R, Boucher L, Heinicke S, Chen D, Stark C,

Breitkreutz A, Kolas N, O'Donnell L, Reguly T, Nixon J, Ramage L, Winter A, Sellam A, Chang C, Hirschman J, Theesfeld C, Rust J, Livstone MS, Dolinski K, Tyers M (2015) The BioGRID interaction database: 2015 update. Nucleic Acids Res 43(Database issue):D470–D478. doi:10.1093/nar/gku1204

16. Szklarczyk D, Franceschini A, Wyder S, Forslund K, Heller D, Huerta-Cepas J, Simonovic M, Roth A, Santos A, Tsafou KP, Kuhn M, Bork P, Jensen LJ, von Mering C (2015) STRING v10: protein-protein interaction networks, integrated over the tree of life. Nucleic Acids Res 43(Database issue):D447–D452. doi:10.1093/nar/gku1003
17. Salwinski L, Miller CS, Smith AJ, Pettit FK, Bowie JU, Eisenberg D (2004) The database of interacting proteins: 2004 update. Nucleic Acids Res 32(suppl 1):D449–D451. doi:10.1093/nar/gkh086
18. Bader GD, Betel D, Hogue CW (2003) BIND: the biomolecular interaction network database. Nucleic Acids Res 31(1):248–250
19. Chatr-aryamontri A, Ceol A, Palazzi LM, Nardelli G, Schneider MV, Castagnoli L, Cesareni G (2007) MINT: the molecular INTeraction database. Nucleic Acids Res 35(suppl 1):D572–D574. doi:10.1093/nar/gkl950
20. Orchard S, Ammari M, Aranda B, Breuza L, Briganti L, Broackes-Carter F, Campbell NH, Chavali G, Chen C, del Toro N, Duesbury M, Dumousseau M, Galeota E, Hinz U, Iannuccelli M, Jagannathan S, Jimenez R, Khadake J, Lagreid A, Licata L, Lovering RC, Meldal B, Melidoni AN, Milagros M, Peluso D, Perfetto L, Porras P, Raghunath A, Ricard-Blum S, Roechert B, Stutz A, Tognolli M, van Roey K, Cesareni G, Hermjakob H (2014) The MIntAct project—IntAct as a common curation platform for 11 molecular interaction databases. Nucleic Acids Res 42(D1):D358–D363. doi:10.1093/nar/gkt1115
21. Shannon P, Markiel A, Ozier O, Baliga NS, Wang JT, Ramage D, Amin N, Schwikowski B, Ideker T (2003) Cytoscape: a software environment for integrated models of biomolecular interaction networks. Genome Res 13(11):2498–2504
22. Kohl M, Wiese S, Warscheid B (2011) Cytoscape: software for visualization and analysis of biological networks. Methods Mol Biol 696:291–303
23. Barabasi A-L, Oltvai ZN (2004) Network biology: understanding the cell's functional organization. Nat Rev Genet 5(2):101–113
24. Ho Y, Gruhler A, Heilbut A, Bader GD, Moore L, Adams S-L, Millar A, Taylor P, Bennett K, Boutilier K, Yang L, Wolting C, Donaldson I, Schandorff S, Shewnarane J, Vo M, Taggart J, Goudreault M, Muskat B, Alfarano C, Dewar D, Lin Z, Michalickova K, Willems AR, Sassi H, Nielsen PA, Rasmussen KJ, Andersen JR, Johansen LE, Hansen LH, Jespersen H, Podtelejnikov A, Nielsen E, Crawford J, Poulsen V, Sorensen BD, Matthiesen J, Hendrickson RC, Gleeson F, Pawson T, Moran MF, Durocher D, Mann M, Hogue CWV, Figeys D, Tyers M (2002) Systematic identification of protein complexes in Saccharomyces cerevisiae by mass spectrometry. Nature 415(6868):180–183
25. Ge H (2000) UPA, a universal protein array system for quantitative detection of protein–protein, protein–DNA, protein–RNA and protein–ligand interactions. Nucleic Acids Res 28(2):e3
26. Keene JD, Komisarow JM, Friedersdorf MB (2006) RIP-Chip: the isolation and identification of mRNAs, microRNAs and protein components of ribonucleoprotein complexes from cell extracts. Nat Protoc 1(1):302–307
27. Uetz P, Giot L, Cagney G, Mansfield TA, Judson RS, Knight JR, Lockshon D, Narayan V, Srinivasan M, Pochart P, Qureshi-Emili A, Li Y, Godwin B, Conover D, Kalbfleisch T, Vijayadamodar G, Yang M, Johnston M, Fields S, Rothberg JM (2000) A comprehensive analysis of protein-protein interactions in Saccharomyces cerevisiae. Nature 403(6770):623–627
28. Zahiri J, Bozorgmehr JH, Masoudi-Nejad A (2013) Computational prediction of protein–protein interaction networks: algorithms and resources. Curr Genomics 14(6):397–414. doi:10.2174/1389202911314060004
29. Pan A, Lahiri C, Rajendiran A, Shanmugham B (2015) Computational analysis of protein interaction networks for infectious diseases. Brief Bioinform. doi:10.1093/bib/bbv059
30. Jeong H, Mason SP, Barabasi AL, Oltvai ZN (2001) Lethality and centrality in protein networks. Nature 411(6833):41–42
31. Hartwell LH, Hopfield JJ, Leibler S, Murray AW (1999) From molecular to modular cell biology. Nature 402(6761 Suppl):47–52
32. Berenstein AJ, Piñero J, Furlong LI, Chernomoretz A (2015) Mining the modular structure of protein interaction networks. PLoS One 10(4), e0122477. doi:10.1371/journal.pone.0122477
33. Saito R, Smoot ME, Ono K, Ruscheinski J, Wang P-L, Lotia S, Pico AR, Bader GD, Ideker T (2012) A travel guide to Cytoscape plugins. Nat Methods 9(11):1069–1076. doi:10.1038/nmeth.2212

34. Han K, Park B, Kim H, Hong J, Park J (2004) HPID: The human protein interaction database. Bioinformatics 20(15):2466–2470. doi:10.1093/bioinformatics/bth253
35. Chen JY, Mamidipalli S, Huan T (2009) HAPPI: an online database of comprehensive human annotated and predicted protein interactions. BMC Genomics 10(Suppl 1):S16
36. Pratt D, Chen J, Welker D, Rivas R, Pillich R, Rynkov V, Ono K, Miello C, Hicks L, Szalma S, Stojmirovic A, Dobrin R, Braxenthaler M, Kuentzer J, Demchak B, Ideker T (2015) NDEx, the network data exchange. Cell Syst 1(4):302–305. doi:10.1016/j.cels.2015.10.001
37. Cowley MJ, Pinese M, Kassahn KS, Waddell N, Pearson JV, Grimmond SM, Biankin AV, Hautaniemi S, Wu J (2012) PINA v2.0: mining interactome modules. Nucleic Acids Res 40(D1):D862–D865. doi:10.1093/nar/gkr967
38. Chisanga D, Keerthikumar S, Pathan M, Ariyaratne D, Kalra H, Boukouris S, Mathew NA, Saffar HA, Gangoda L, Ang C-S, Sieber OM, Mariadason JM, Dasgupta R, Chilamkurti N, Mathivanan S (2016) Colorectal cancer atlas: an integrative resource for genomic and proteomic annotations from colorectal cancer cell lines and tissues. Nucleic Acids Res 44(D1):D969–D974. doi:10.1093/nar/gkv1097
39. Breitkreutz B-J, Stark C, Tyers M (2003) Osprey: a network visualization system. Genome Biol 4(3):R22

SPRINGER LICENSE
TERMS AND CONDITIONS

Feb 19, 2017

This Agreement between David Chisanga ("You") and Springer ("Springer") consists of your license details and the terms and conditions provided by Springer and Copyright Clearance Center.

License Number	4052720950784
License date	Feb 19, 2017
Licensed Content Publisher	Springer
Licensed Content Publication	Springer eBook
Licensed Content Title	Network Tools for the Analysis of Proteomic Data
Licensed Content Author	David Chisanga
Licensed Content Date	Jan 1, 2017
Type of Use	Thesis/Dissertation
Portion	Full text
Number of copies	1
Author of this Springer article	Yes and you are the sole author of the new work
Order reference number	
Title of your thesis / dissertation	Protein Networks in Cancer Cells
Expected completion date	Aug 2017
Estimated size(pages)	300
Requestor Location	David Chisanga LIMS Level 5,La Trobe University Kingsbury Drive Melbourne, Victoria 3083 Australia Attn: David Chisanga
Billing Type	Invoice
Billing Address	David Chisanga LIMS Level 5,La Trobe University Kingsbury Drive Melbourne, Australia 3083 Attn: David Chisanga
Total	0.00 AUD
Terms and Conditions	

Introduction
The publisher for this copyrighted material is Springer. By clicking "accept" in connection with completing this licensing transaction, you agree that the following terms and conditions apply to this transaction (along with the Billing and Payment terms and conditions established by Copyright Clearance Center, Inc. ("CCC"), at the time that you opened your Rightslink account and that are available at any time at http://myaccount.copyright.com).
Limited License
With reference to your request to reuse material on which Springer controls the copyright, permission is granted for the use indicated in your enquiry under the following conditions:
- Licenses are for one-time use only with a maximum distribution equal to the number stated in your request.
- Springer material represents original material which does not carry references to other sources. If the material in question appears with a credit to another source, this permission is not valid and authorization has to be obtained from the original copyright holder.

Print This Page

Nucleic Acids Research, 2016, Vol. 44, Database issue **D969–D974**
doi: 10.1093/nar/gkv1097

Colorectal cancer atlas: An integrative resource for genomic and proteomic annotations from colorectal cancer cell lines and tissues

David Chisanga[1], Shivakumar Keerthikumar[2], Mohashin Pathan[2], Dinuka Ariyaratne[2], Hina Kalra[2], Stephanie Boukouris[2], Nidhi Abraham Mathew[2], Haidar Al Saffar[2], Lahiru Gangoda[2], Ching-Seng Ang[3], Oliver M. Sieber[4,5], John M. Mariadason[6,7,8], Ramanuj Dasgupta[9], Naveen Chilamkurti[1] and Suresh Mathivanan[2,*]

[1]Department of Computer Science and Information Technology, La Trobe University, Bundoora, Victoria 3086, Australia, [2]Department of Biochemistry and Genetics, La Trobe Institute for Molecular Science, La Trobe University, Melbourne, Victoria 3086, Australia, [3]The Bio21 Molecular Science and Biotechnology Institute, University of Melbourne, Parkville, Victoria 3010, Australia, [4]Walter and Eliza Hall Institute of Medical Research, Parkville, Victoria 3052, Australia, [5]Faculty of Medicine, Dentistry and Health Sciences, Department of Medical Biology, University of Melbourne, Parkville, Victoria 3052, Australia, [6]Olivia Newton John Cancer Research Institute, Melbourne, Victoria 3084, Australia, [7]Ludwig Institute for Cancer Research, Melbourne-Austin Branch, Victoria 3084, Australia, [8]School of Cancer Medicine, La Trobe University, Melbourne, Victoria 3084, Australia and [9]Genome Institute of Singapore, A*STAR, 60 Biopolis Street, Singapore 138672, Singapore

Received August 15, 2015; Revised October 02, 2015; Accepted October 11, 2015

ABSTRACT

In order to advance our understanding of colorectal cancer (CRC) development and progression, biomedical researchers have generated large amounts of OMICS data from CRC patient samples and representative cell lines. However, these data are deposited in various repositories or in supplementary tables. A database which integrates data from heterogeneous resources and enables analysis of the multidimensional data sets, specifically pertaining to CRC is currently lacking. Here, we have developed Colorectal Cancer Atlas (http://www.colonatlas.org), an integrated web-based resource that catalogues the genomic and proteomic annotations identified in CRC tissues and cell lines. The data catalogued to-date include sequence variations as well as quantitative and non-quantitative protein expression data. The database enables the analysis of these data in the context of signaling pathways, protein–protein interactions, Gene Ontology terms, protein domains and post-translational modifications. Currently, Colorectal Cancer Atlas contains data for >13 711 CRC tissues, >165 CRC cell lines, 62 251 protein identifications, >8.3 million MS/MS spectra, >18 410 genes with sequence variations (404 278 entries) and 351 pathways with sequence variants. Overall, Colorectal Cancer Atlas has been designed to serve as a central resource to facilitate research in CRC.

INTRODUCTION

Colorectal cancer (CRC) is the third most common form of cancer and has the fourth highest mortality rate in the world (1). In order to advance our understanding of the initiation and progression of this disease, biomedical researchers have performed global analyses of the genome, epigenome, transcriptome, proteome and metabolome of CRC patient samples and representative cell lines (2–5). According to The Cancer Genome Atlas Network (3), APC, TP53, KRAS, PIK3CA, FBXW7, SMAD4, TCF7L2 and NRAS are the most frequently mutated genes in CRC. Identification of these mutations and associated pathways has advanced our understanding of CRC, is enabling the sub-classification of this disease and is unveiling potential new avenues for treatment.

Due to the significant advancements in high-throughput technologies, vast amounts of multidimensional data relevant to the biology of CRC have been generated. To extract meaningful biological insights from these data, researchers previously needed to collate data from a large number of studies. To facilitate this process, a series of databases have been created. For example, cancer gene mutations are currently catalogued in databases including TCGA (3), COS-

---

[*] To whom correspondence should be addressed. Tel: +61 03 9479 2565; Fax: +61 03 9479 1226; Email: S.Mathivanan@latrobe.edu.au

Table 1. Colorectal cancer atlas statistics

Protein entries	62 251
MS/MS spectra	8 378 422
Primary tissues	13 711
Cell lines	165
Genes with sequence variants	18 410
Gene sequence variants	404 278
Pathways with genes having sequence variants	351
Pathways with genes having no sequence variants	1657
Cell lines with drug sensitivity	27
PTMs	88 819
PTMs affected by sequence variants	1631
Protein–protein interactions	253 700

MIC (6), TumorPortal (7), IntOGen (8), Network of Cancer Genes (9) and TSGene (10). These databases provide valuable information of gene variations for a number of tumor types including CRC, however, they are not specifically designed to integrate sequence variations with proteomic data. NetGestal (11) is a web-based framework that allows for integration of OMIC data from multiple species in the context of biological networks (12) and contains data pertaining to human CRC from TCGA. However, there is currently no user-friendly online resource specifically pertaining to CRC which catalogues genomic and proteomic data from literature, databases and TCGA, integrates the sequence variations with protein domain, post-translational modifications and protein–protein interactions.

Here, we describe Colorectal Cancer Atlas (http://www.colonatlas.org), an integrated web-based resource which catalogues genomic and proteomic data from CRC tissues and cell lines. Data catalogued include; quantitative and non-quantitative protein expression, sequence variations, cellular signaling pathways, protein–protein interactions, Gene Ontology terms, protein domains and post-translational modifications (PTMs). Data pertaining to genomic sequence variations and protein expression have been manually curated from the scientific literature and collated from other publicly available databases. Colorectal Cancer Atlas is designed to enable a user to search for a specific mutation in any particular cell line, and search for cell lines with and without specific mutations. Currently, Colorectal Cancer Atlas contains data for >13 711 primary CRC tissues, >165 CRC cell lines, 62 251 protein identifications, >8.3 million MS/MS spectra, >18 410 genes with sequence variations, 404 278 sequence variation entries, 351 pathways with sequence variants, 88 819 PTMs and 253 700 protein–protein interactions (Table 1).

DATABASE ARCHITECTURE AND WEB INTERFACE

Colorectal Cancer Atlas is a web-based application developed using Zope2 (version 2.8.7–1), a python-based web framework. The back end database is MySQL (version 5.0.95), a well-established open source database. The web pages were developed using Hyper Text Markup Language (HTML) in combination with JavaScript for front end functionality, while Python (version 2.4.3), a scripting language was used for database connectivity. JavaScript modules include DataTables (version 1.10.4) for the development of interactive data tables, Data-Driven Documents (D3JS) for the development of interactive protein–protein interaction networks, and Highcharts (version 4.1.6) for the development of interactive heat maps and column charts.

GENOMIC DATA SETS

Colorectal Cancer Atlas catalogues gene sequence variations present in primary CRC tissues and cell lines which were collated by manual curation of the scientific literature. In addition, the database contains genomic variations identified in CRC cell lines sequenced in-house. For cell lines, where available, the gender and age of the patient is provided, along with the specific cell type, doubling time, culture properties and stage of cancer. This information was obtained from the Cancer Cell Line Encyclopedia (13), ATCC (http://www.atcc.org), COSMIC database and literature. Sequence variation details including the type of sequence variants, putative mutational effects, nucleotide change and amino acid changes are displayed.

PROTEOMIC DATA SETS

Colorectal Cancer Atlas also catalogues proteomic data collated from multiple resources including the scientific literature (e.g. Zhang *et al.* (5)), Human Protein Atlas (14), Human Proteinpedia (15) and Human Protein Reference Database (16). Experimental techniques used in generating these data included mass spectrometry, Western blotting, immunohistochemistry, confocal microscopy, immunoelectron microscopy and fluorescence-activated cell sorting (FACS). In addition, publicly available label-free quantitative mass spectrometry data for CRC cell lines and tissues were re-analyzed using an in-house proteomics pipeline in order to provide standardized data. The proteomics pipeline involved conversion of raw mass spectrometry data files into the Mascot Generic File Format (MGF) using MsConvert with peak picking (17). The MGF files were then searched using X! Tandem (Sledgehammer edition version 2013.09.01.1) (18) against a target and decoy Human RefSeq protein database. Peptides were further filtered using <5% false discovery rate (FDR) as a cut-off, and quantified using the Normalized Spectral Abundance Factor (NSAF) method (19).

COLORECTAL CANCER ATLAS PROVIDES AN INTEGRATED VIEW OF MULTIPLE DATA TYPES

Colorectal Cancer Atlas provides an integrated view of the sequence variations and the proteomic data. Mass spectrometry-based quantitative proteomic data are depicted as heat maps and column charts in the respective molecular pages (Figure 1), and users are able to filter the data sets based on the FDR. The database also contains protein expression data generated using immunohistochemistry, Western blotting, FACS, confocal and immunoelectron microscopy. The database also includes protein data derived from various cellular fractions including the nucleus, cytoplasm, membrane, the secretome (20) and exosomes (21) (from ExoCarta (22)).

The integration of sequence variants with proteomic data is designed to facilitate the prediction of functional effects of the protein. For each gene, Colorectal Cancer Atlas

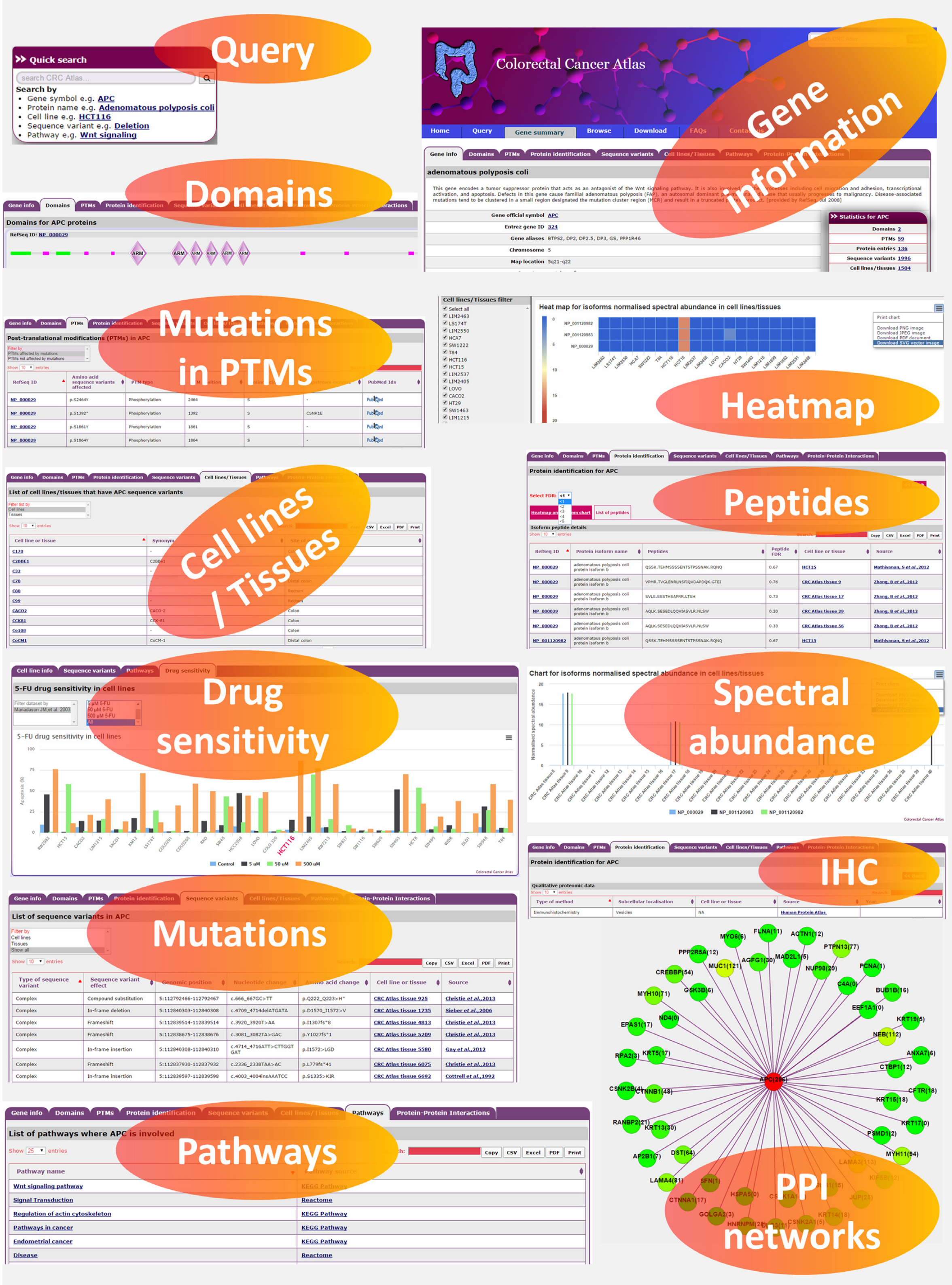
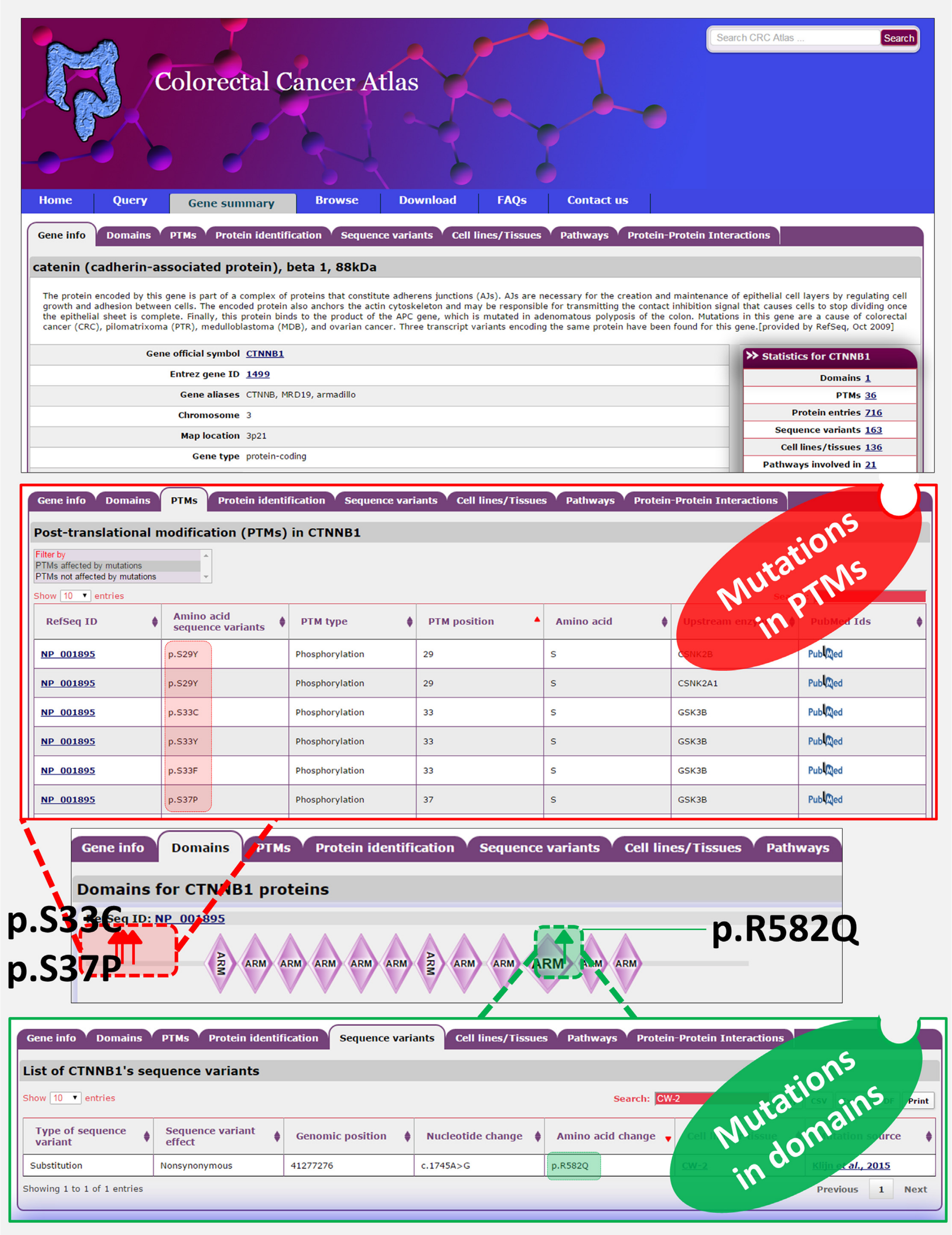


Figure 1. Snapshot of Colorectal Cancer Atlas features. An overview of proteomic and genomic data features for APC gene is displayed. A user can query the database using a gene symbol or a protein name. A gene information page will provide the users with details pertaining to protein domains, post-translational modifications (PTM), reported mutations in cell lines/tissues, quantitative protein expression, pathway, protein–protein interaction (PPI) and cell line drug sensitivity.



catenin (cadherin-associated protein), beta 1, 88kDa

The protein encoded by this gene is part of a complex of proteins that constitute adherens junctions (AJs). AJs are necessary for the creation and maintenance of epithelial cell layers by regulating cell growth and adhesion between cells. The encoded protein also anchors the actin cytoskeleton and may be responsible for transmitting the contact inhibition signal that causes cells to stop dividing once the epithelial sheet is complete. Finally, this protein binds to the product of the APC gene, which is mutated in adenomatous polyposis of the colon. Mutations in this gene are a cause of colorectal cancer (CRC), pilomatrixoma (PTR), medulloblastoma (MDB), and ovarian cancer. Three transcript variants encoding the same protein have been found for this gene.[provided by RefSeq, Oct 2009]

Gene official symbol	CTNNB1
Entrez gene ID	1499
Gene aliases	CTNNB, MRD19, armadillo
Chromosome	3
Map location	3p21
Gene type	protein-coding

Statistics for CTNNB1

Domains	1
PTMs	36
Protein entries	716
Sequence variants	163
Cell lines/tissues	136
Pathways involved in	21

Post-translational modification (PTMs) in CTNNB1

Filter by
PTMs affected by mutations
PTMs not affected by mutations

Show 10 entries

RefSeq ID	Amino acid sequence variants	PTM type	PTM position	Amino acid	Upstream enzy	PubMed Ids
NP_001895	p.S29Y	Phosphorylation	29	S	CSNK2B	PubMed
NP_001895	p.S29Y	Phosphorylation	29	S	CSNK2A1	PubMed
NP_001895	p.S33C	Phosphorylation	33	S	GSK3B	PubMed
NP_001895	p.S33Y	Phosphorylation	33	S	GSK3B	PubMed
NP_001895	p.S33F	Phosphorylation	33	S	GSK3B	PubMed
NP_001895	p.S37P	Phosphorylation	37	S	GSK3B	PubMed

Domains for CTNNB1 proteins

RefSeq ID: NP_001895

List of CTNNB1's sequence variants

Show 10 entries

Search: CW-2

Type of sequence variant	Sequence variant effect	Genomic position	Nucleotide change	Amino acid change	Cell li[illegible] tissue	[illegible]tation source
Substitution	Nonsynonymous	41277276	c.1745A>G	p.R582Q	CW-2	Klijn et al., 2015

Showing 1 to 1 of 1 entries

Figure 2. PTMs and domains in β-catenin are affected due to mutation. Snapshot of β-catenin molecular page is displayed. The PTMs affected by mutations can be viewed in the tab PTMs. Mutations in β-catenin at positions important for phosphorylation (S33, S37, T41 and S45) allows for the stabilization of β-catenin and constitutive activation of the Wnt signaling pathway. The upstream kinases responsible for the phosphorylation is also provided along with the literature reference. Likewise, mutations in the armadillo domain can be viewed by correlating the sequence variants and the domain span regions. For example, mutations in the armadillo domain (p.R582) in β-catenin have been described which have been reported to alter the binding of β-catenin to TCF4 (24).

enables parallel visualization of CRC-associated sequence variants with quantitative protein expression across CRC cell lines and tissues. In addition, PTMs, and protein domains affected by the sequence variation can be visualized (Figure 1), enabling the potential effect of sequence variants on protein function to be easily ascertained. For example, β-catenin mutations in positions S33, S37, T41 and S45 occur in CRC, all of which are critical for phosphorylation (23). Mutations in these serine/threonine residues allow for the stabilization of β-catenin and constitutive activation of the Wnt signaling pathway. Similarly, Colorectal Cancer Atlas displays sequence variations in known protein domains which can provide valuable insight into the putative effect on protein function. For example, mutations in the armadillo domain (R582) in β-catenin have been described which have been reported to alter the binding of β-catenin to TCF4 (24) (Figure 2).

Colorectal Cancer Atlas also provides a graphical representation of known protein interactions (obtained from BioGrid (25) and Human Protein Resource Database (16)), where each protein is depicted as a node with a specific colour and intensity corresponding to the number of sequence variants in the encoding gene (Figure 1). Furthermore, Colorectal Cancer Atlas integrates biological pathways with gene sequence variants. Biological Pathways were obtained from Reactome (26), KEGG (27), Cell map and HumanCyc. For example, as shown in Figure 1, sequence variants in APC are implicated in dysregulation of the Wnt signaling pathway and actin cytoskeletal remodeling. Finally, Colorectal Cancer Atlas contains data on 5-flurouracil (5-FU) drug sensitivity for CRC cell lines curated from the literature (studies using at least three CRC cell lines (28)). Users can view the sensitivity profile of a cell line of interest relative to other CRC cells.

ACCESSING COLORECTAL CANCER ATLAS

Users can search Colorectal Cancer Atlas through the home, query or browse pages (Supplementary Figure S1). In addition, the website features a navigation menu and a search box at the top of the page. The database can be queried by gene symbol, Entrez Gene ID, protein name, cell line name or pathway. The browse page provides users with the option to access the database by categorized lists of genes, sequence variations, cell lines and techniques. The browse page allows the users to search for sequence variations in genes of interest and displays them in interactive color-coded table format. The gene information page includes gene details, associated GO terms, sequence variations (displayed in an interactive table), domain details, PTMs, a protein data page leading to experimental techniques and quantitative data with an interactive heat map, a column chart for spectral abundance and a list of detected peptides. Other information includes a list of cell lines and tissues that contain sequence variants in a given gene, a list of pathways in which the gene is involved, and an interactive protein–protein interaction network for the protein encoded by the gene. The cell line page provides details of the cell line, an interactive table of gene sequence variants identified in the cell line, an interactive table of dysregulated pathways and 5-FU drug sensitivity profile. Data curated in Colorectal Cancer Atlas are available as tab-delimited files and is free for download to all users. Using the custom database option, the tab delimited data can also be uploaded into FunRich (29), a functional enrichment analysis tool to identify classes of genes/proteins that are overrepresented in a specific category.

FUTURE DIRECTIONS

Colorectal Cancer Atlas will be continuously updated with more studies as they become available and additional features. Studies currently being curated include Wnt signaling activity determined by the TOPFLASH assay, and genomic and proteomic data generated from patient derived xenografts.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

FUNDING

Australian NH&MRC fellowship [1016599 to S.M.] and Ramaciotti Establishment grant. The funders had no role in study design, data collection and analysis, decision to publish or preparation of the manuscript. Funding for open access charge: Australian NH&MRC fellowship [1016599] and Ramaciotti Establishment grant.

Conflict of interest statement. None declared.

REFERENCES

1. Jemal,A., Bray,F., Center,M.M., Ferlay,J., Ward,E. and Forman,D. (2011) Global cancer statistics. *CA Cancer J. Clin.*, **61**, 69–90.
2. Atlas,T.C.G. (2012) Comprehensive molecular characterization of human colon and rectal cancer. *Nature*, **487**, 330–337.
3. Cancer Genome Atlas Network. (2012) Comprehensive molecular characterization of human colon and rectal cancer. *Nature*, **487**, 330–337.
4. Sadanandam,A., Lyssiotis,C.A., Homicsko,K., Collisson,E.A., Gibb,W.J., Wullschleger,S., Ostos,L.C., Lannon,W.A., Grotzinger,C., Del Rio,M. *et al.* (2013) A colorectal cancer classification system that associates cellular phenotype and responses to therapy. *Nat. Med.*, **19**, 619–625.
5. Zhang,B., Wang,J., Wang,X., Zhu,J., Liu,Q., Shi,Z., Chambers,M.C., Zimmerman,L.J., Shaddox,K.F., Kim,S. *et al.* (2014) Proteogenomic characterization of human colon and rectal cancer. *Nature*, **513**, 382–387.
6. Forbes,S.A., Beare,D., Gunasekaran,P., Leung,K., Bindal,N., Boutselakis,H., Ding,M., Bamford,S., Cole,C., Ward,S. *et al.* (2015) COSMIC: exploring the world's knowledge of somatic mutations in human cancer. *Nucleic Acids Res.*, **43**, D805–D811.
7. Lawrence,M.S., Stojanov,P., Mermel,C.H., Robinson,J.T., Garraway,L.A., Golub,T.R., Meyerson,M., Gabriel,S.B., Lander,E.S. and Getz,G. (2014) Discovery and saturation analysis of cancer genes across 21 tumour types. *Nature*, **505**, 495–501.
8. Gundem,G., Perez-Llamas,C., Jene-Sanz,A., Kedzierska,A., Islam,A., Deu-Pons,J., Furney,S.J. and Lopez-Bigas,N. (2010) IntOGen: integration and data mining of multidimensional oncogenomic data. *Nat. Methods*, **7**, 92–93.
9. An,O., Pendino,V., D'Antonio,M., Ratti,E., Gentilini,M. and Ciccarelli,F.D. (2014) NCG 4.0: the network of cancer genes in the era of massive mutational screenings of cancer genomes. *Database*, **2014**, doi:10.1093/database/bau015.
10. Zhao,M., Sun,J. and Zhao,Z. (2013) TSGene: a web resource for tumor suppressor genes. *Nucleic Acids Res.*, **41**, D970–D976.

11. Shi,Z., Wang,J. and Zhang,B. (2013) NetGestalt: integrating multidimensional omics data over biological networks. *Nat. Methods*, **10**, 597–598.
12. Zhu,J., Shi,Z., Wang,J. and Zhang,B. (2015) Empowering biologists with multi-omics data: colorectal cancer as a paradigm. *Bioinformatics*, **31**, 1436–1443.
13. Barretina,J., Caponigro,G., Stransky,N., Venkatesan,K., Margolin,A.A., Kim,S., Wilson,C.J., Lehar,J., Kryukov,G.V., Sonkin,D. *et al.* (2012) The Cancer Cell Line Encyclopedia enables predictive modelling of anticancer drug sensitivity. *Nature*, **483**, 603–307.
14. Uhlen,M., Fagerberg,L., Hallstrom,B.M., Lindskog,C., Oksvold,P., Mardinoglu,A., Sivertsson,A., Kampf,C., Sjostedt,E., Asplund,A. *et al.* (2015) Proteomics. Tissue-based map of the human proteome. *Science*, **347**, 1260419.
15. Mathivanan,S., Ahmed,M., Ahn,N.G., Alexandre,H., Amanchy,R., Andrews,P.C., Bader,J.S., Balgley,B.M., Bantscheff,M., Bennett,K.L. *et al.* (2008) Human Proteinpedia enables sharing of human protein data. *Nat. Biotechnol.*, **26**, 164–167.
16. Keshava Prasad,T.S., Goel,R., Kandasamy,K., Keerthikumar,S., Kumar,S., Mathivanan,S., Telikicherla,D., Raju,R., Shafreen,B., Venugopal,A. *et al.* (2009) Human protein reference database–2009 update. *Nucleic Acids Res.*, **37**, D767–D772.
17. Chambers,M.C., Maclean,B., Burke,R., Amodei,D., Ruderman,D.L., Neumann,S., Gatto,L., Fischer,B., Pratt,B., Egertson,J. *et al.* (2012) A cross-platform toolkit for mass spectrometry and proteomics. *Nat. Biotechnol.*, **30**, 918–920.
18. Craig,R. and Beavis,R.C. (2004) TANDEM: matching proteins with tandem mass spectra. *Bioinformatics*, **20**, 1466–1467.
19. Paoletti,A.C., Parmely,T.J., Tomomori-Sato,C., Sato,S., Zhu,D., Conaway,R.C., Conaway,J.W., Florens,L. and Washburn,M.P. (2006) Quantitative proteomic analysis of distinct mammalian Mediator complexes using normalized spectral abundance factors. *Proc. Natl. Acad. Sci. U.S.A.*, **103**, 18928–18933.
20. Mathivanan,S., Ji,H., Tauro,B.J., Chen,Y.S. and Simpson,R.J. (2012) Identifying mutated proteins secreted by colon cancer cell lines using mass spectrometry. *J. Proteomics*, **76**, 141–149.
21. Keerthikumar,S., Gangoda,L., Liem,M., Fonseka,P., Atukorala,I., Ozcitti,C., Mechler,A., Adda,C.G., Ang,C.S. and Mathivanan,S. (2015) Proteogenomic analysis reveals exosomes are more oncogenic than ectosomes. *Oncotarget*, **6**, 15375–15396.
22. Keerthikumar,S., Chisanga,D., Ariyaratne,D., Al Saffar,H., Anand,S., Zhao,K., Samuel,M., Pathan,M., Jois,M., Chilamkurti,N. *et al.* (2015) ExoCarta: a web-based compendium of exosomal cargo. *J. Mol. Biol.*, doi:10.1016/j.jmb.2015.09.019.
23. Wang,Z., Vogelstein,B. and Kinzler,K.W. (2003) Phosphorylation of beta-catenin at S33, S37, or T41 can occur in the absence of phosphorylation at T45 in colon cancer cells. *Cancer Res.*, **63**, 5234–5235.
24. Fasolini,M., Wu,X., Flocco,M., Trosset,J.Y., Oppermann,U. and Knapp,S. (2003) Hot spots in Tcf4 for the interaction with beta-catenin. *J. Biol. Chem.*, **278**, 21092–21098.
25. Chatr-aryamontri,A., Breitkreutz,B.-J., Oughtred,R., Boucher,L., Heinicke,S., Chen,D., Stark,C., Breitkreutz,A., Kolas,N., O'Donnell,L. *et al.* (2015) The BioGRID interaction database: 2015 update. *Nucleic Acids Res.*, **43**, D470–D478.
26. Milacic,M., Haw,R., Rothfels,K., Wu,G., Croft,D., Hermjakob,H., D'Eustachio,P. and Stein,L. (2012) Annotating Cancer Variants and Anti-Cancer Therapeutics in Reactome. *Cancers*, **4**, 1180.
27. Kanehisa,M., Goto,S., Sato,Y., Kawashima,M., Furumichi,M. and Tanabe,M. (2014) Data, information, knowledge and principle: back to metabolism in KEGG. *Nucleic Acids Res.*, **42**, D199–D205.
28. Mariadason,J.M., Arango,D., Shi,Q., Wilson,A.J., Corner,G.A., Nicholas,C., Aranes,M.J., Lesser,M., Schwartz,E.L. and Augenlicht,L.H. (2003) Gene expression profiling-based prediction of response of colon carcinoma cells to 5-fluorouracil and camptothecin. *Cancer Res.*, **63**, 8791–8812.
29. Pathan,M., Keerthikumar,S., Ang,C.S., Gangoda,L., Quek,C.Y., Williamson,N.A., Mouradov,D., Sieber,O.M., Simpson,R.J., Salim,A. *et al.* (2015) FunRich: An open access standalone functional enrichment and interaction network analysis tool. *Proteomics*, **15**, 2597–2601.

OXFORD UNIVERSITY PRESS LICENSE
TERMS AND CONDITIONS

Feb 19, 2017

This Agreement between David Chisanga ("You") and Oxford University Press ("Oxford University Press") consists of your license details and the terms and conditions provided by Oxford University Press and Copyright Clearance Center.

License Number	4052720358017
License date	
Licensed content publisher	Oxford University Press
Licensed content publication	Nucleic Acids Research
Licensed content title	Colorectal cancer atlas: An integrative resource for genomic and proteomic annotations from colorectal cancer cell lines and tissues
Licensed content author	Chisanga, David; Keerthikumar, Shivakumar
Licensed content date	2015-10-22
Type of Use	Thesis/Dissertation
Institution name	
Title of your work	Protein Networks in Cancer Cells
Publisher of your work	n/a
Expected publication date	Aug 2017
Permissions cost	0.00 AUD
Value added tax	0.00 AUD
Total	0.00 AUD
Requestor Location	David Chisanga LIMS Level 5,La Trobe University Kingsbury Drive Melbourne, Victoria 3083 Australia Attn: David Chisanga
Publisher Tax ID	GB125506730
Billing Type	Invoice
Billing Address	David Chisanga LIMS Level 5,La Trobe University Kingsbury Drive Melbourne, Australia 3083 Attn: David Chisanga
Total	0.00 AUD
Terms and Conditions	

STANDARD TERMS AND CONDITIONS FOR REPRODUCTION OF MATERIAL FROM AN OXFORD UNIVERSITY PRESS JOURNAL

1. Use of the material is restricted to the type of use specified in your order details.
2. This permission covers the use of the material in the English language in the following territory: world. If you have requested additional permission to translate this material, the terms and conditions of this reuse will be set out in clause 12.
3. This permission is limited to the particular use authorized in (1) above and does not allow you to sanction its use elsewhere in any other format other than specified above, nor does it

apply to quotations, images, artistic works etc that have been reproduced from other sources which may be part of the material to be used.

4. No alteration, omission or addition is made to the material without our written consent. Permission must be re-cleared with Oxford University Press if/when you decide to reprint.

5. The following credit line appears wherever the material is used: author, title, journal, year, volume, issue number, pagination, by permission of Oxford University Press or the sponsoring society if the journal is a society journal. Where a journal is being published on behalf of a learned society, the details of that society must be included in the credit line.

6. For the reproduction of a full article from an Oxford University Press journal for whatever purpose, the corresponding author of the material concerned should be informed of the proposed use. Contact details for the corresponding authors of all Oxford University Press journal contact can be found alongside either the abstract or full text of the article concerned, accessible from www.oxfordjournals.org Should there be a problem clearing these rights, please contact journals.permissions@oup.com

7. If the credit line or acknowledgement in our publication indicates that any of the figures, images or photos was reproduced, drawn or modified from an earlier source it will be necessary for you to clear this permission with the original publisher as well. If this permission has not been obtained, please note that this material cannot be included in your publication/photocopies.

8. While you may exercise the rights licensed immediately upon issuance of the license at the end of the licensing process for the transaction, provided that you have disclosed complete and accurate details of your proposed use, no license is finally effective unless and until full payment is received from you (either by Oxford University Press or by Copyright Clearance Center (CCC)) as provided in CCC's Billing and Payment terms and conditions. If full payment is not received on a timely basis, then any license preliminarily granted shall be deemed automatically revoked and shall be void as if never granted. Further, in the event that you breach any of these terms and conditions or any of CCC's Billing and Payment terms and conditions, the license is automatically revoked and shall be void as if never granted. Use of materials as described in a revoked license, as well as any use of the materials beyond the scope of an unrevoked license, may constitute copyright infringement and Oxford University Press reserves the right to take any and all action to protect its copyright in the materials.

9. This license is personal to you and may not be sublicensed, assigned or transferred by you to any other person without Oxford University Press’s written permission.

10. Oxford University Press reserves all rights not specifically granted in the combination of (i) the license details provided by you and accepted in the course of this licensing transaction, (ii) these terms and conditions and (iii) CCC’s Billing and Payment terms and conditions.

11. You hereby indemnify and agree to hold harmless Oxford University Press and CCC, and their respective officers, directors, employs and agents, from and against any and all claims arising out of your use of the licensed material other than as specifically authorized pursuant to this license.

12. Other Terms and Conditions:

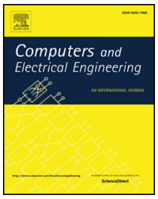
v1.4

Computers and Electrical Engineering 67 (2018) 267–277

Contents lists available at ScienceDirect

Computers and Electrical Engineering

journal homepage: www.elsevier.com/locate/compeleceng



Integration of heterogeneous ‘omics’ data using semi-supervised network labelling to identify essential genes in colorectal cancer[☆]

David Chisanga[a], Shivakumar Keerthikumar[b], Suresh Mathivanan[b], Naveen Chilamkurti[a,*]

[a] *Department of Computer Science and Information Technology, La Trobe University, Bundoora, Victoria 3086, Australia*
[b] *Department of Biochemistry and Genetics, La Trobe Institute for Molecular Science, La Trobe University, Melbourne, Victoria 3086, Australia*

ARTICLE INFO

Keywords:
Proteomics
Protein-protein interactions
Big data
Genomics
Networks
Colorectal cancer
Biomarkers
Machine learning

ABSTRACT

Colorectal cancer (CRC) is the third most common form of cancer and has the fourth highest mortality rate in the world. To understand the origin and progression of this disease, biomedical researchers undertake global analyses of omics data of CRC patient samples and representative cell lines. However, due to the heterogeneity and high dimensionality nature of `omics’ data, traditional tools for analysing this sort of data are inadequate and the heterogeneous nature of cancer makes the process of identifying essential genes very difficult. ‘Omics’ is a term that is used to refer to areas of study in biology that end with the ending ‘omics’ such as genomics, proteomics and metabolomics. This paper uses network theory-based methods to address the problem of high dimensionality in omics datasets and applies network propagation to address the problem of heterogeneity in both omics datasets and cancer in identifying the essential genes. The method successfully identifies known essential genes in CRC as well as a new set of genes that are likely to be essential in the study of CRC.

1. Introduction

Network theory, the study of how complex systems interact is widely applied in fields such as computer networks, social networks, and interactome networks in systems biology [1]. Network metrics such as node degree are often used to prioritise nodes within a network. Similarly, one of the main goals in cancer research is the identification of biomarkers or essential genes that can be used to understand the development or progression of a specific cancer type such as Colorectal cancer (CRC).

To prioritise these genes, researchers often study the complex interactions between the numerous molecules within cells such as proteins, deoxyribonucleic acid (DNA), ribonucleic acid (RNA) and other small molecules. The molecules are obtained from the global profiling of patient samples as well as representative cell lines at multiple layers, these layers constitute what is today referred to as ‘omics’ data. ‘Omics’ is an informal term that is used to refer to areas of study in biology that end with the term ‘omics’ such as genomics, proteomics and metabolomics [2]. The interactions, on the other hand, are collectively known as interactome networks and provide a global picture of how molecular interactions influence cellular behaviour, an example being protein-protein interactions (PPI) [3].

Omics data is highly dimensional in nature, coupled with this, is the heterogeneity of cancer whereby two individuals with the same type of cancer may have a different set of biomarkers. This makes identifying and prioritising cancer-related genes a challenging

---

[☆] Reviews processed and recommended for publication to the Editor-in-Chief by Guest Editor Dr. A. Sangaiah.

[*] Corresponding author at: Department of Computer Science and Information Technology, La Trobe University, Bundoora, Victoria 3086, Australia.
E-mail address: N.Chilamkurti@latrobe.edu.au (N. Chilamkurti).

https://doi.org/10.1016/j.compeleceng.2018.03.039
Received 4 July 2017; Received in revised form 19 March 2018; Accepted 19 March 2018

and daunting task that cannot be achieved using traditional statistical methods. As such, network theory provides a means by which complexity in such instances can be used to model the cellular system behaviour. Barabási, et al. [4] provides a summation of the application of network-based metrics in associating omics-related molecules to disease. Other works in [5–7] applied network-based methods in areas such as identifying and associating genes to disease as well as identifying drug targets in various cancer types. In [8,9], integrated network-based methods with machine learning techniques are applied in reducing the dimensionality of omics data and building models to predict genes associated with the disease as well as classify multiple cancer types. While the integration of omics data with networks has been gaining momentum over the years, a typical recurring theme in most of the research has been the use of a single type of omics data as opposed to integrating the various types of omics data which are heterogeneous in nature.

In this paper, we used an integrated approach to identify essential genes in colorectal cancer, a type of cancer that originates in the bowel, is the third most common form of cancer and has the fourth highest cancer mortality rate in the world [10]. The integrated approach employed a semi-supervised learning algorithm to propagate heterogeneous omics data into a protein-protein interaction network, which was followed by a downstream enrichment analysis to validate and understand the role of the predicted potential essential genes in CRC.

The rest of the paper is organised as follows: Section 2 provides a description of the materials and methods used as well as an overview of related works, Section 3 provides a discussion of the experimental results and the implications of the findings. The paper concludes with a summary of the findings and the future directions of the research.

2. Materials and methods

2.1. Proteomics data

We used proteomics and genomics data as the input to our method. Proteomics data consisted of protein-protein interactions. Weighted protein-protein interactions were downloaded from HIPPIE Version 2.0 [11], an online web-based database resource for weighted protein-protein interactions. The weights in the interactions show the confidence in the interaction between two proteins and are calculated by the authors based on the amount and reliability of evidence supporting an interaction. The protein-protein interaction dataset was then filtered to leave out interactions with a confidence score of 0 after which 16,728 number of unique proteins and 276, 183 number of interactions remain. These were then assembled into a network using NetworkX, a Python package for network manipulation and analysis.

2.2. Genomics data

Genomics data comprised gene somatic mutations and gene differential expression status for CRC patients and representative cell lines. Previously, we collated genomics data related to CRC into a web-based resource called the Colorectal Cancer Atlas [12]. It is this data together with The Cancer Genome Atlas (TCGA) patient data obtained from COSMIC [13] that we used as the genomics input data to our method. Using the corresponding genes for the proteins identified above, we obtained gene mutation details of 564 CRC patients from TCGA.

From the mutation dataset, we then filtered out all silent mutations and for each gene with a mutation in each sample, we represented its status using a binary number (1 if a mutation was present and 0 if not present) regardless of the number of mutations for a gene in each sample. The mutation data were then represented as a matrix, $\mathbf{M}$ (16,728 × 564) with rows representing genes and columns representing a gene's mutation station status in each sample. The same was repeated for gene differential expression status in TCGA patient data. This was then represented as a matrix, $\mathbf{D}$ (16,728 × 564) with rows representing genes and columns representing the differential expression status of genes in each sample. The gene differential expression status was denoted 1 for under-regulated or up-regulated genes and 0 for genes not differentially expressed.

2.3. Theory/calculation

To identify essential genes, we use a method that integrates the different datasets discussed in the materials and methods section. Fig. 1 provides a summary of the approach taken in this paper.

2.4. Disease gene prioritisation using network theory methods

A network or a graph is defined as a set of objects (nodes) linked together by lines (edges) [1]. A network is, therefore, represented as an ordered pair G = (V, E) where V is the set of nodes and E is the set of edges. By grouping a collection of objects as a set of nodes and using edges to represent relationships between these objects, researchers have used networks to reduce the complexity of large systems. Molecular networks in biology provide a global representation of the complex interactions between various molecules within a cell such as DNA, RNA and other small molecules.

When it comes to disease-gene prioritisation, many researchers use networks to associate genes with diseases. A naïve approach that is usually taken is to predict those genes that have neighbours associated with a disease as being more likely to be implicated in such a disease, that is using the concept of “guilty by association”. Such methods that implicate neighbours as having the likelihood of being associated with a disease include node degree as well as shortest path methods. However, these methods are prone to false positives because of the biases that exist in current molecular networks’ datasets where proteins which are well studied tend to have

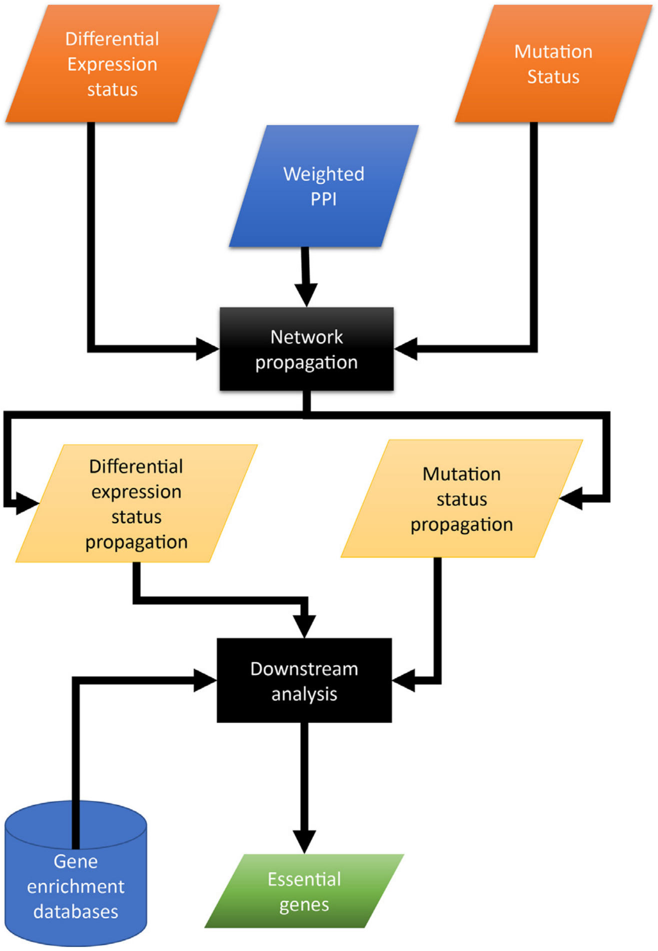

Differential Expression status
Mutation Status
Weighted PPI
Network propagation
Differential expression status propagation
Mutation status propagation
Downstream analysis
Gene enrichment databases
Essential genes

Fig. 1. Architecture of model. Differential expression status and mutation propagation status data were propagated through the network. The propagation results were then integrated together to form the features which were used in the further downstream analysis.

more interactions than those that are not. In addition, biological networks tend to obey the concept of the “small world” property where each node is reachable to another node through a series of links with other nodes and as such, the average number of hops needed to get to the furthest node from any given node is small [14].

2.5. Network propagation

Here, we used network propagation, a semi-supervised labelling algorithm first proposed by Zhou et al. [15] and further extended by Vanunu et al. [16] and Ruffalo et al. [17]. The objective was to determine the extent to which a gene's mutation status or differential expression status is propagated globally in a network, ultimately affecting the topology of the network. The propagation results were then used to perform enrichment analysis to validate and determine roles played by the predicted essential genes in CRC. The input to the algorithm was a semi-labelled vector of gene mutation status M_v or differential expression status D_v, and a protein-protein interaction network as shown in Eq. (1);

$$G(V, E, w) \tag{1}$$

where V is the set of proteins, E is the set of interactions and w is the set of interaction confidence scores (weight). The aim was to be able to determine the distance of the proteins in V (those that have not been labelled as either mutated or differentially expressed) from those that have been labelled as either mutated or differentially expressed.

For each node $v \epsilon V$, we let N (v) be indicative of the direct neighbours of v in G. Let F: $V \rightarrow \Re$ be the propagation function where F (v) denotes the distance of a protein from those that are either differentially expressed or mutated as shown in Eq. (2). Let Y: $V \rightarrow [0,1]$ denote some prior knowledge function matching genes known to be differentially expressed or mutated as one (1) and zero (0) if not.

$$F(v) = \alpha\left[\sum_{\mu \in N(v)} F(u)w'(v, u)\right] + (1 - \alpha)Y(v) \tag{2}$$

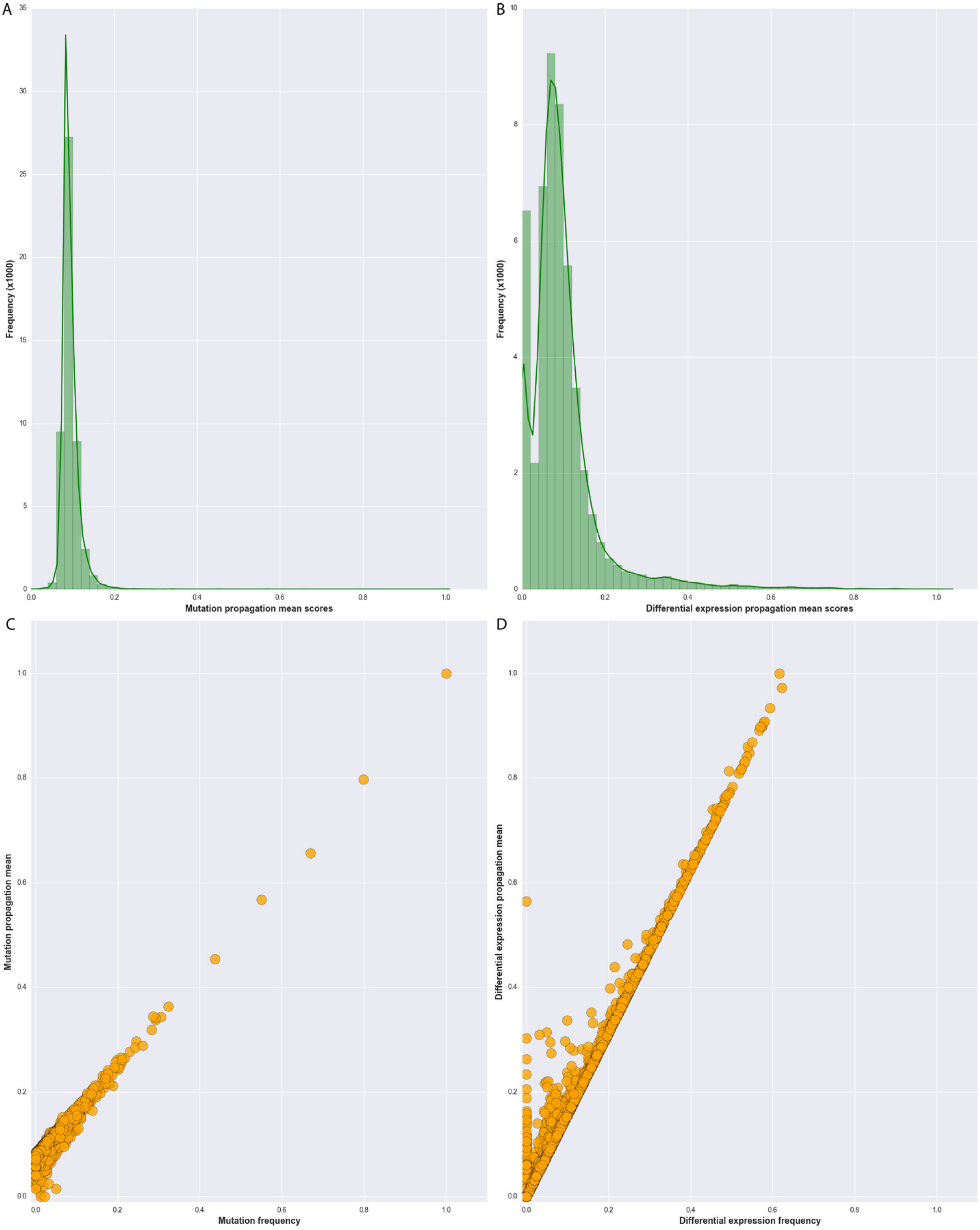


Fig. 2. Summary of propagation scores in TCGA samples. (a) distribution of mutation propagation scores with the scores indicating the proximity of the genes to the genes that had a mutation (b) differential expression status propagation scores where the scores indicate the proximity of genes to genes that were differentially expressed between normal and tumour samples (c) the relationship between the mean of mutation propagation scores was compared against the mutation frequency to understand the relationship between the two. The results showed that genes with a high mutation frequency had high propagation scores, these were filtered and only those with lower scores were obtained for further analysis (d) the relationship between the mean of differential expression status propagation scores was compared against differential expression frequency and is in (c) only those genes with lower propagation scores were obtained for further downstream analysis.

Table 1
Network propagation algorithm sensitivity scores. The sensitivity scores are used to measure the consistency of network propagation in labelling correctly known genes as having high propagation scores similar to their previous labels.

	Mutation	Differential expression
Number of correct label	104,505	565,582
Number of incorrect label	5	0
Sensitivity	0.999 ≈ 1	1.0

where w' is a [v]x[v] matrix and is a Laplacian normalised form of w as described below, the parameter $\alpha \in (0, 1)$ weighs the relative importance of the two constraints discussed above, F and Y are vectors of size [n] where Y is the prior knowledge. We used an iterative procedure to compute network propagation as in Eq. (3):

$$F^t = \alpha W' F^{t-1} + (1 - \alpha)Y \tag{3}$$

where $F^1 = Y$ and W' represents w'. The iterative algorithm can be described as a process whereby proteins for which prior genomic (mutated or differentially expressed) information exists iteratively pass on this information to their neighbouring nodes and every other node further propagates the information from the previous round to its neighbours repeatedly until convergence.

W' is a square matrix which represents the Laplacian normalisation of an [n]x[n] adjacency matrix W which is built from the set of confidence scores between interactions. We built an adjacency matrix W, with a non-zero indicating an interaction between the two nodes and vice-versa. We then use Laplacian normalisation to get the matrix W' as shown in Eq. (4):

$$W' = D^{-1/2} W D^{-1/2} \tag{4}$$

where $D^{-1/2}$ is a diagonal matrix such that D (i, i) is the sum of row i of W.

After computation of the normalised weighted matrix W', for each sample in our data sets, we then iteratively computed the propagation scores for each of the nodes in the PPI by setting Y as the prior knowledge vector where all the nodes whose corresponding genes known to either be mutated or differentially expressed were set to 1 and 0 otherwise. The propagation was computed separately by propagating node mutation status using the mutation status dataset as well as for the differential expression status dataset resulting in P_m for mutation based propagation scores and P_d for differentially expressed based propagation scores. The propagation scores are then used to perform the following computations; propagation mean scores for genes across the sample, standard deviation, covariance which is then used to perform further downstream analysis to identify essential genes.

3. Results and discussion

3.1. Propagation of omics data

Network propagation of mutation status and that of differential expression status data is performed, Fig. 2 shows the distribution of scores in TCGA samples respectively. The figure also shows the relationships between the propagation scores against their corresponding status data. From this, it is shown that genes with a high-frequency rate of mutation or differential expression across samples are labelled with a propagation score close to their initial label in the prior knowledge dataset. This is further confirmed by the sensitivity of the algorithm as shown in Table 1. The sensitivity is calculated by comparing the total number of correctly predicted/labelled genes against the total number of genes known a priori.

We hypothesise that genes with high mutation or differential propagation scores have a closer relationship to those genes that are either mutated or differentially expressed while those with low propagation scores are distant from the mutated or differentially expressed genes in the network. Based on the remaining filtered genes, we then pick the genes with propagation scores and perform enrichment analysis.

3.2. Enrichment analysis of mutation status propagation scores

To obtain an understanding of the relevance of the propagation results to CRC, we performed enrichment analysis on the propagation results using FunRich [18]. In Fig. 3(a) and (b), enrichment analysis of the genes with high mean mutation status propagation scores reveal that these genes are highly enriched in several cancers in the COSMIC database, furthermore, of these, it is found that 47 are also part of the COSMIC cancer gene census, as shown in Table A1.

In addition, we also performed the biological process and molecular function enrichment to determine processes and functions most likely to be affected by the genes with high mutation status propagation scores as shown in Fig. 3(c) and (d) respectively. Of interest to us from the biological processes were homophilic cell adhesion and cell adhesion, as in [19] it is shown that these two processes play an important role in contact inhibition. Contact inhibition is cellular changes that lead to the termination of cell migration and proliferation because of signals transduced when one cell comes into physical contact with another cell. Nonetheless, in tumour microenvironments, it is shown that contact inhibition is lost due to the molecular changes in cell-cell adhesion, this, in turn, leads to cell proliferation and/or migration. This, therefore, means that changes in cell adhesion properties in cancer micro tumour environment play a key role in cancer progression and metastasis [20]. Genes enriched in the two pathways are also shown in Table A2.

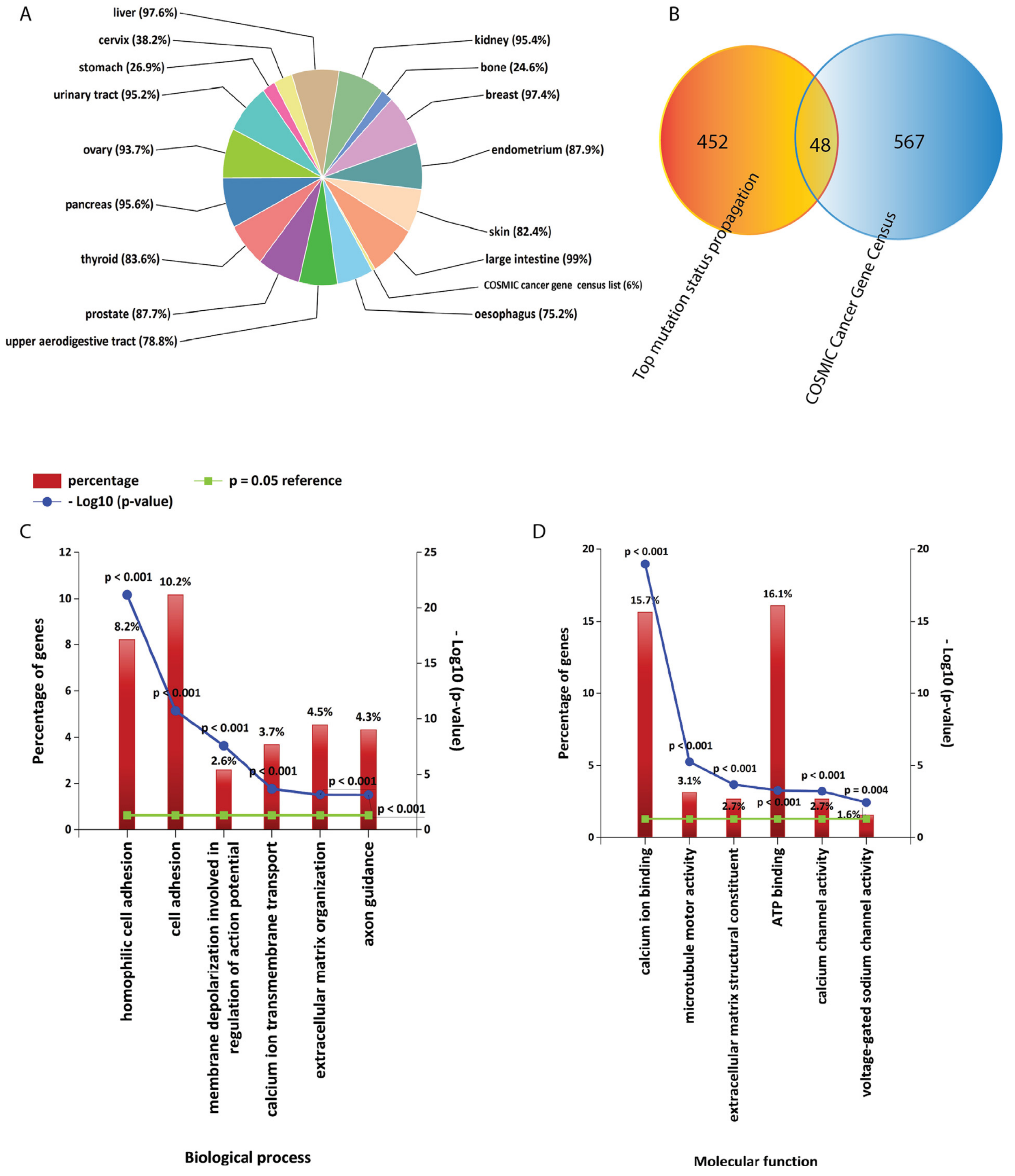


Fig. 3. Enrichment analysis of genes with high mutation status propagation scores. (a) shows that genes with high mutation status propagation scores are highly enriched in different types of cancers in COSMIC, (b) shows that 47 genes short-listed from the high mutation propagation scores are also found in the COSMIC census gene lists, (c) shows the biological process of the genes with high mutation status propagation scores, and (d) shows the molecular function enrichment of genes high propagation scores.

On the other hand, from the molecular function enrichment, it was found that genes that had high mutation propagation scores were also enriched in calcium ion binding and ATP binding molecular functions as shown in Table A3. Calcium ion binding is part of the calcium cell signalling pathways whereby proteins bind to the Ca^{2+} ion. This pathway is important in regulating various cellular processes. A dysregulation of calcium ion binding function in cancer cells has been linked to the hyperpolarization of tumour cells

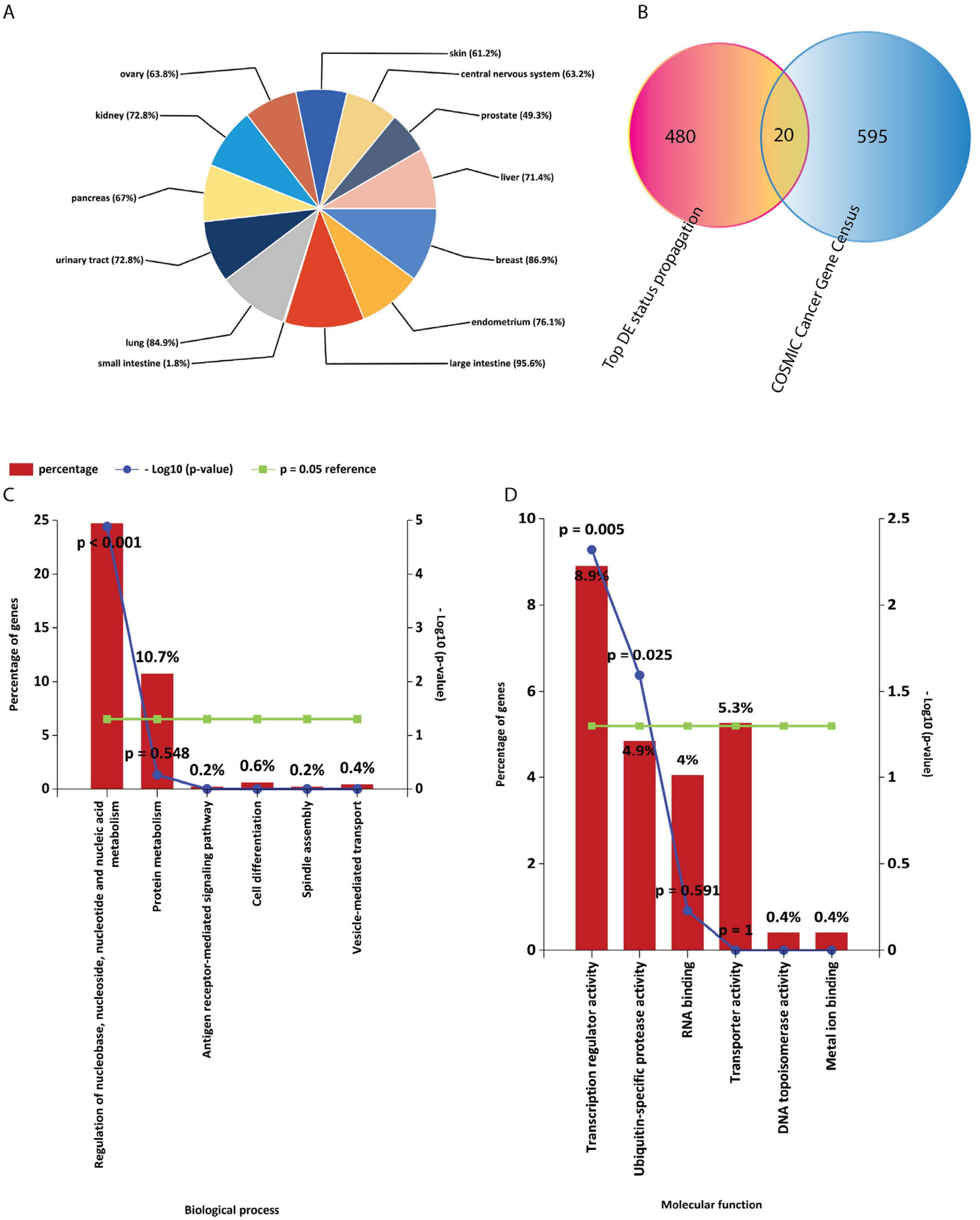


Fig. 4. Enrichment analysis of genes with high differential expression status propagation scores; (a) shows that genes with high differential expression status propagation scores are highly enriched in various forms of cancers in COSMIC, (b) shows that 20 genes short-listed from the high differential expression scores are also found in the COSMIC census gene lists, (c) shows that genes are only significantly enriched in one biological process, and (d) shows that genes with high differential scores are only enriched in two molecular functions.

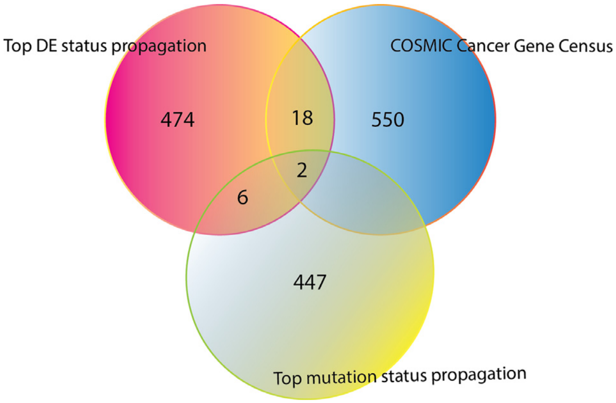

Top DE status propagation
COSMIC Cancer Gene Census
474
18
550
2
6
447
Top mutation status propagation

Fig. 5. The Venn diagram shows the genes found to be closer to genes that are differentially expressed and have a mutation. The Venn diagram also shows the genes that are found on the COSMIC's cancer gene census.

and impacts cancer cell proliferation and metastasis [21]. In addition, related to calcium ion binding functionality is the ATP (adenosine triphosphate) binding function which acts as a source of energy needed by the ATP-binding cassette transporters to translocate substrates across membranes. The increased expression of ATP-binding cassette members has been shown to play a role in multi-drug resistance in diseases such as cancer [22]. These results, therefore, demonstrate that by propagating mutation status across the network we can prioritise high scoring genes and their associated pathways and processes that are most likely to be affected by mutated counterparts.

3.3. Enrichment analysis of differential expression status propagation scores

We also performed enrichment analysis for genes with high mean differential expression status propagation scores as shown in Fig. 4, and Tables A1, A4 and A5. The results show that similar to the mutation status propagation enrichment previously discussed, genes with high differential expression status propagation scores are highly enriched in various types of cancer from the COSMIC database. A comparison against COSMIC's cancer gene census shows that 20 of these genes are also found on the census list and the biological process and molecular function enrichments are not as significant as above. Nonetheless, of the significantly enriched molecular functions, dysregulation in Ubiquitin-specific protease activity has been shown to be associated with cancer and members have been studied as potential drug targets in the treatment of cancer [23].

3.4. Linking mutation status and differential expression status scores

From the two lists of genes with high propagation scores, we filter for genes that appear in both lists obtaining a set of 8 genes as shown in Fig. 5, two of which are also enriched in COSMIC cancer gene census. These genes are considered as being close to both mutated and differentially expressed genes in the network. The following is the list of the identified genes; RALY, **ASXL1**, DIDO1, AP11A, ZC3H13, UGGT2, CCAR2 and **SMAD4**. ASXL1 and SMAD4 are known to be driver genes in cancer and are part of the COSMIC cancer gene census dataset. For instance, ASXL1 has been implicated in myelodysplastic syndrome (MDS) and chronic myelomonocytic leukaemia (CML) while SMAD4 has been implicated in the following cancer types; colorectal, pancreatic, and small intestine. On the other hand, a literature search of the remaining six genes shows that they have also been implicated in some of form of cancer with varying roles ranging from resistance, metastasis and cell proliferation. For example, RALY is a gene that codes for the protein RNA-binding protein and in [24] has been implicated to play a role in the development of drug resistance in CRC, DIDO1 is a gene which codes for the protein death inducer-obliterator and is involved in apoptosis or cell death and has been found to affect cell viability and anchorage in CRC cells and CCAR2 has been implicated in other forms of cancer [25].

3.5. Conclusions and future directions

The rate at which omics data is generated has over the years been rising substantially and is expected to rise further due to the continued decline in the cost and the advancements in high-throughput technologies such as next-generation sequencing technologies. As such traditional statistical methods can no longer be relied upon as a way of analysing such gigantic amounts of data. Network analysis, the evaluation of how nodes relate to one another coupled with new machine learning methods, has over the years become an integral tool for analysing high throughput data such as omics data.

In this paper, we demonstrated how heterogeneous omics datasets can be integrated by use of network-based methods and how features can be prioritised using a semi-supervised technique coupled with further downstream analysis. We found that the method successfully identified the essential genes in CRC. Further, we also identified new genes that may play a role CRC in the development and progression of cancer.

However, the genes that were predicted in this paper need further experimental validation to understand their specific roles in

CRC. In addition, this study was limited by the lack of vast amounts of paired wild-type and mutant data, this, in turn, made it difficult to further explore our findings and incorporate soft computing techniques. Future works include fine-tuning the current model and validating the predicted genes using wet laboratory experiments. We also plan on incorporating new machine learning techniques such as deep learning using neural networks.

Acknowledgements

SM is supported by the Australian NHMRC fellowship (1016599) and Ramaciotti Establishment grant. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

This research was supported by the use of the NeCTAR Research Cloud, a collaborative Australian research platform supported by the National Collaborative Research Infrastructure Strategy.

Appendices

Table A1

Genes found in COSMIC cancer gene census from propagation scores.

Genes from mutation propagation scores	Genes from differential expression propagation scores
AKAP9; ARID1A; ASXL1; ATM; ATP2B3; ATRX; BCL9L; BCORL1; BRAF; CASC5; CHD4; CIITA; FAT1; FAT4; FBXW7; GNAS; HLA-A; MT2A; KMT2D; KRAS; LIFR; LRP1B; MED12; MN1; MTOR; MYH11; NCOR2; NF1; NRAS; NRG1; PBRM1; PDE4DIP; PIK3CA; POLE; PREX2; PTPRT; RBM15; RNF213; RNF43; ROS1; RUNX1T1; SALL4; SMAD4; SPECC1; TCF7L2; TPR; ZFHX3	ASXL1; CUX1; ERCC5; MAP2K4; MYC; NONO; PHF6; PLCG1; RAD21; RB1; SMAD2; SMAD4; SRC; SS18; SS18L1; STAG2; TFE3; TOP1; UBR5; ZMYM

Table A2

Biological process enrichment of genes with high mutation propagation scores.

Biological process	Enriched genes

Homophilic cell adhesion	FAT3; ROBO2; FAT4; SDK1; ROBO1; DCHS2; PCDHA12;
	DSCAM; PCDHA7; PTPRT; PCDH10; FAT1; PCDHA6; TENM3; CELSR1; DCHS1; PCDH9; PCDH11X; CELSR2; CDH18; PCDHB3; PCDH20; PCDHB8; PCDHA3; SDK2; CDH23; PCDHA11; PCDH17; PCDHA2; PCDHA9; PCD-HGB2; PCDHA5; DSCAML1; PCDHGA11; PCDHA4; PCDHA10;

Homophilic cell adhesion	FAT3; ROBO2; FAT4; SDK1; ROBO1; DCHS2; PCDHA12;
	DSCAM; PCDHA7; PTPRT; PCDH10; FAT1; PCDHA6; TENM3; CELSR1; DCHS1; PCDH9; PCDH11X; CELSR2; CDH18; PCDHB3; PCDH20; PCDHB8; PCDHA3; SDK2; CDH23; PCDHA11; PCDH17; PCDHA2; PCDHA9; PCD-HGB2; PCDHA5; DSCAML1; PCDHGA11; PCDHA4; PCDHA10;

Table A3

Molecular function enrichment of genes with high mutation propagation scores.

Molecular function	Enriched genes
Calcium ion binding	PROC; TTN; FAT3; PCLO; DST; CACNA1B; NRXN1; FAT4; RYR2; FLG; DCHS2; PCDHA12; MEGF8; CACNA1E; FBN2; TENM2; CDHA7; LRP1B; BRAF; TCHH; ADGRL3; RYR1; PCDH10; GPR98; FAT1; PCDHA6; SLIT3; HMCN1; RYR3; CELSR1; SPTA1; CUBN; FBN3; DCHS1; PCDH9; PCDH11X; CELSR2; CDH18; FBN1; VCAN; PCDHB3; TBC1D9; DNAH7; HRNR; MEGF6; TPO; PCDH20; SLC25A12; PCDHB8; SLC25A23; CDHA3; CDH23; PCDHA11; PKDREJ; PCDH17; PCDHA2; PCDHA9; LTBP3; PCDHGB2; LRP2; PCDHA5; STAB1; PCDHGA11; EFCAB6; ITPR1; ASTN2; LTBP4; PCDHA4; TNNC1; FSTL5; PLCH2; PCDHA10; MATN4;

ATP binding	TTN; PIK3CA; ABCA13; CACNA1B; OBSCN; DNAH10; DNAH14; DNAH2; KIF26B; ABCA7; CHD4; BRAF; ATP10A; RYR1; HELZ2; ATRX; DNAH5; DNAH9; MYH11; NLRP7; MDN1; DNAH8; EP400; LATS2; NAV3; TTBK1; MYH13; MYO18B; DNAH1; ACACB; ATM; DNAH11; ATP2B4; DNA2; SPEG; MYO3A; EPHB1; NWD1; SRCAP; DNAH7; ATP8B2; PHA3; ADCY8; WNK1; NLRP4; KIF1A; CIITA; CHD6; KIF4B; ATP13A3; ATP2B3; ROS1; NLRX1; SETX; ATP7A;
	SCN8A; LRRK2; DNAH6; ATP8B1; ABCA4; SMARCA2; DNAH3; ABCA12; MYO15A; NLRP5; MTOR; ATP11A; SMC1B; TTLL11; EPHA10; NRK; MYH3;

Table A4
Biological process enrichment of genes with high differential expression propagation scores.

Biological process	Enriched genes
Regulation of nucleobase, nucleoside, nucleotide and nucleic acid metabolism	DDX27; NELFCD; TAF4; NCOA6; GMEB2; TCFL5; RPRD1B; PLAGL2; RBM39; RALY; DHX35; CSTF1; PCIF1; NCOA5; SUPT20H; HNF4A; RNF6; ASXL1; RNF113A; PHF20; ADNP; PDRG1; ZMYND8; MRGBP; TGIF2; GTF2F2; JADE3; NUFIP1; ZGPAT; ZFP64; FTSJ1; GZF1; PAN3; NONO; PQBP1; PARP4; UCKL1; SAP18; DKC1; UPF3A; NKRF; GTF3A; XRN2; PHF8; HNRNPH2; PABPC1; TRMT2B; ZNF696; HSF1; WBP4; ERCC5; ZNF623; CHRAC1; MYBL2; MAF1; ZNF34; PRICKLE3; ZHX3; RAD21; ZBTB33; TOP1MT; FAM50A; POLA1; UTP14C; TAF2; DCAF13; ZNF7; CRNKL1; TFDP1; DIS3; MED30; GTF2E2; RBM41; HUWE1; CNOT7; ZNF217; TOP1; UPF3B; TDRD3; MORF4L2; V39H1; CTPS2; GRHL2; HDAC6; PDS5B; HDAC8; PUF60; ZNF706; SCML2; ZFP41; INTS6; DSCC1; RBMX2; ZNF41; ZC3H13; SS18; DNMT3B; TFE3; POLR3D; HMGB1; PHF6; E2F1; POLR1D; KRBOX4; ASH2L; RB1; ZNF335; MBD1; CUX1; THOC2; ZNF337; CBFA2T2; SMAD4; MECP2; MYC; ID1; ZMYM2; ZSCAN25; ZMIZ2; ZC3H3; ZNF24; ZNF250;

Table A5
Molecular function enrichment of genes with high differential expression propagation scores.

Molecular function	Enriched genes
Transcription regulator activity	NELFCD; NCOA6; PLAGL2; RBM39; PCIF1; NCOA5; SUPT20H; HNF4A; RNF6; ASXL1; PDRG1; ZMYND8; MRGBP; SS18L1; JADE3; PQBP1; SAP18; SCAND1; UXT; NKRF; ZNF696; MAF1; PRICKLE3; ZHX3; ZBTB33; MED30; SMAD2; CNOT7; MORF4L2; HDAC6; HDAC8; SCML2; ZC3H13; SS18; PHF6; RB1; ZNF335; MBD1; CUX1; ID1; ZMYM2; ZMIZ2; ZNF24; ZNF250;
Ubiquitin- specific pro- tease activity	CUL4A; RNF114; LNX2; ITCH; NEURL2; UBE2C; RNF219; TMEM189; COPS5; UBR5; UCHL3; UBL3; FBXL3; CUL1; UBL4A; SUGT1; UBE2D4; USP12; UBE2A; PSMD10; RNF216; PJA1; SCRIB; UBE2G1;

Supplementary materials

Supplementary material associated with this article can be found, in the online version, at doi:10.1016/j.compeleceng.2018.03.039.

References

[1] Chisanga D, Keerthikumar S, Chilamkurti N. Network tools for the analysis of proteomic data. In: Mathivanan S, editor. Proteome bioinformatics. Springer; 2017.

[2] Horgan RP, Kenny LC. 'Omic' technologies: genomics, transcriptomics, proteomics and metabolomics. Obstetric Gynaecol 2011;13(3):189–95.

[3] Chatr-aryamontri A, Oughtred R, Boucher L, Rust J, Chang C, Kolas NK, O'Donnell L, Oster S, Theesfeld C, Sellam A, Stark C, Breitkreutz B-J, Dolinski K, Tyers M. The BioGRID interaction database: 2017 update. Nucleic Acids Res 2017;45(D1):D369–79.

[4] Barabási A-L, Gulbahce N, Loscalzo J. Network medicine: a network-based approach to human disease. Nat Rev Genet 2011;12(1):56–68.

[5] Chen Y, Xu R. Network-based gene prediction for plasmodium falciparum malaria towards genetics-based drug discovery. BMC Genom 2015;16(7):S9.

[6] Peng J, Bai K, Shang X, Wang G, Xue H, Jin S, Cheng L, Wang Y, Chen J. Predicting disease-related genes using integrated biomedical networks. BMC Genom 2017;18(1):1043.

[7] Cui Y, Cai M, Stanley HE. Discovering disease-associated genes in weighted protein-protein interaction networks. Physica A 2017.

[8] Gui T, Dong X, Li R, Li Y, Wang Z. Identification of hepatocellular carcinoma-related genes with a machine learning and network analysis (in eng). J Comput Biol 2015;22(1):63–71.

[9] Philips S, Wu HY, Li L. Using machine learning algorithms to identify genes essential for cell survival (in eng). BMC Bioinformatics 2017;18(Suppl 11):397.

[10] World Cancer Research Fund International. (2017). Colorectal Cancer Statistics. Available: http://www.wcrf.org/int/cancer-facts-figures/data-specific-cancers/colorectal-cancer-statistics.

[11] Alanis-Lobato G, Andrade-Navarro MA, Schaefer MH. HIPPIE v2.0: enhancing meaningfulness and reliability of protein–protein interaction networks. Nucleic Acids Res 2017;45(D1):D408–14.

[12] Chisanga D, Keerthikumar S, Pathan M, Ariyaratne D, Kalra H, Boukouris S, Mathew NA, Saffar HA, Gangoda L, Ang C-S, Sieber OM, Mariadason JM, Dasgupta R, Chilamkurti N, Mathivanan S. Colorectal cancer atlas: An integrative resource for genomic and proteomic annotations from colorectal cancer cell lines and tissues. Nucleic Acids Res 2016;44(D1):D969–74.

[13] Forbes SA, Beare D, Boutselakis H, Bamford S, Bindal N, Tate J, Cole CG, Ward S, Dawson E, Ponting L, Stefancsik R, Harsha B, Kok CY, Jia M, Jubb H, Sondka Z, Thompson S, De T, Campbell PJ. COSMIC: somatic cancer genetics at high-resolution. Nucleic Acids Res 2017;45(D1):D777–83.

[14] Cowen L, Ideker T, Raphael BJ, Sharan R. Network propagation: a universal amplifier of genetic associations (in eng). Nat Rev Genet 2017;18(9):551–62.

[15] Zhou D, Bousquet O, Lal TN, Weston J, Schölkopf B. Learning with local and global consistency. Advances in neural information processing systems. 2004. p. 321–8.

[16] Vanunu O, Magger O, Ruppin E, Shlomi T, Sharan R. Associating genes and protein complexes with disease via network propagation. PLoS Comput Biol 2010;6(1):e1000641.

[17] Ruffalo M, Koyutürk M, Sharan R. Network-based integration of disparate omic data to identify" silent players" in cancer. PLoS Comput Biol 2015;11(12):e1004595.

[18] Pathan M, Keerthikumar S, Chisanga D, Alessandro R, Ang C-S, Askenase P, Batagov AO, Benito-Martin A, Camussi G, Clayton A, Collino F, Di Vizio D, Falcon-Perez JM, Fonseca P, Fonseka P, Fontana S, Gho YS, Hendrix A, Hoen EN-t, Iraci N, Kastaniegaard K, Kislinger T, Kowal J, Kurochkin IV, Leonardi T, Liang Y, Llorente A, Lunavat TR, Maji S, Monteleone F, Øverbye A, Panaretakis T, Patel T, Peinado H, Pluchino S, Principe S, Ronquist G, Royo F, Sahoo S, Spinelli C, Stensballe A, Théry C, van Herwijnen MJC, Wauben M, Welton JL, Zhao K, Mathivanan S. A novel community driven software for functional enrichment analysis of extracellular vesicles data. J Extracell Ves 2017;6(1):1321455.

[19] Craig SE, Brady-Kalnay SM. Cancer cells cut homophilic cell adhesion molecules and run (in eng). Cancer Res 2011;71(2):303–9.

[20] Hanahan D, Weinberg RobertA. Hallmarks of Cancer: The Next Generation. Cell 2011;144(5):646–74.

[21] Bose T, Cieslar-Pobuda A, Wiechec E. Role of ion channels in regulating Ca(2)(+) homeostasis during the interplay between immune and cancer cells (in eng). Cell Death Dis 2015;6:e1648.

[22] Agarwal T, Annamalai N, Maiti TK, Arsad H. Biophysical changes of ATP binding pocket may explain loss of kinase activity in mutant DAPK3 in cancer: a molecular dynamic simulation analysis (in eng). Gene 2016;580(1):17–25.

[23] Davis MI, Pragani R, Fox JT, Shen M, Parmar K, Gaudiano EF, Liu L, Tanega C, McGee L, Hall MD, McKnight C, Shinn P, Nelson H, Chattopadhyay D, D'Andrea AD, Auld DS, DeLucas LJ, Li Z, Boxer MB, Simeonov A. Small molecule inhibition of the ubiquitin-specific protease USP2 accelerates cyclin D1 degradation and leads to cell cycle arrest in colorectal cancer and mantle cell lymphoma models (in eng). J Biol Chem 2016;291(47):24628–40.

[24] Tsofack SP, Garand C, Sereduk C, Chow D, Aziz M, Guay D, Yin HH, Lebel M. NONO and RALY proteins are required for YB-1 oxaliplatin induced resistance in colon adenocarcinoma cell lines (in eng). Mol Cancer 2011;10:145.

[25] Sillars-Hardebol AH, Carvalho B, Tijssen M, Belien JA, de Wit M, Delis-van Diemen PM, Ponten F, van de Wiel MA, Fijneman RJ, Meijer GA. TPX2 and AURKA promote 20q amplicon-driven colorectal adenoma to carcinoma progression (in eng). Gut 2012;61(11):1568–75.

David Chisanga is currently a Ph.D. candidate in the Department of Computer Science and Information Technology, at La Trobe University, Melbourne, VIC, Australia. He received his BSc degree in Computer Science from the University of Zambia in 2010 and MSc. degree in Information Technology Management from Binary University in 2013. His research interests include Bioinformatics, Data Science, and Machine Learning.

Shivakumar Keerthikumar is currently a cancer bioinformatician at Cancer Research Division, Peter MacCallum Cancer Centre, Melbourne, Australia. His current research combines whole-genome, transcriptomics and DNA-methylation to understanding tumour evolution in prostate cancer. He obtained his Ph.D. in Bioinformatics from the Institute of Bioinformatics, Bangalore, India. He did his post-doctoral research in Intellectual Disability (ID) at Centre for molecular and biomolecular informatics (CMBI), Radboud University Nijmegen Medical Centre, Nijmegen, The Netherlands. He also worked as Research officer, at Department of Biochemistry and Genetics, La Trobe University, Melbourne, Australia, in understanding the role of Exosomes in CRC using genomics and proteomics data. His research interests span from functional genomics to systems biology using computational methods.

Suresh Mathivanan is currently Associate Professor and laboratory head, Biochemistry and Genetics, La Trobe University, Melbourne, VIC, Australia. He obtained his Ph.D. degree from Kuvempu University, India and Johns Hopkins University, USA. His current research areas include cancer, chemoresistance, exosomes, extracellular vesicles, tumour microenvironment, proteomics, and bioinformatics.

Naveen Chilamkurti is currently Associate Professor and Cybersecurity Program Coordinator, Computer Science and Information Technology, La Trobe University, Melbourne, VIC, Australia. He obtained his Ph.D. degree from La Trobe University. His current research areas include intelligent transport systems (ITS), Smart grid computing, vehicular communications, Vehicular cloud, Cybersecurity, wireless multimedia, wireless sensor networks, and Mobile security.