Integrative studies to explore key molecular players involved in cervical squamous cell carcinoma

A thesis submitted to the University of Hyderabad, for the award of

Doctor of Philosophy
in
Biotechnology

By

Thippana Mallikarjuna

(Reg. No. 18LTPH02)

Department of Biotechnology & Bioinformatics

School of Life Sciences

University of Hyderabad

(P.O.) Central University, Gachibowli,

Hyderabad – 500 046 Telangana

India

February 2024

University of Hyderabad
(A central University established in 1974 by Act of Parliament)

School of Life Sciences
Department of Biotechnology & Bioinformatics
Hyderabad-500046

प्रतिष्ठित संस्थान
INSTITUTION OF EMINENCE

CERTIFICATE

This is to certify that this thesis entitled "**Integrative studies to explore key molecular players involved in cervical squamous cell carcinoma**" submitted by **Thippana Mallikarjuna**, bearing registration number **18LTPH02** in partial fulfilment of the requirements for award of Doctor of Philosophy in the School of Life Sciences is a bonafide work carried out by him under my supervision and guidance.

This thesis is free from plagiarism and has not been submitted previously in part or in full to this or any other University or Institution for award of any degree or diploma.

Parts of this thesis have been

A. published in the following journals:

1. **Thippana M**, Dwivedi A, Das A, Palanisamy M, Vindal V. Identification of key molecular players and associated pathways in cervical squamous cell carcinoma progression through network analysis. Proteins. 2023; 1 - 15. doi:10.1002/prot.26502 (**part of thesis objective-iii**)
2. Thummadi NB, **Thippana M**, Vindal V, P. M. Prioritizing the candidate genes related to cervical cancer using the moment of inertia tensor. Proteins. 2021;1-9. doi:10.1002/prot.26226. (**part of thesis objective-i**)

B. Presented in the following conferences:

1. ISMB/ECCB 2023 (31st Conference on Intelligent Systems For Molecular Biology and 22nd Annual European Conference on Computational Biology), Centre de Congrès Lyon, France, July 23-27, 2023.
2. GIW XXXI/ISCB-Asia V 2022 (31st International Conference on Genome Informatics and ISCB Asia V 2022), National Cheng – Kung University (NCKU), Tainan, Taiwan, December 12-14, 2022.

3. ASCS2022 Asian Student Council Symposium 2022, December 10-11, 2022.
4. The 20th International Conference on Bioinformatics (InCoB 2021) held at Kunming, Yunnan, China (Virtual mode), November 2021.
5. Conference on "Proteomics in Agricultural and Healthcare" held at School of Life Sciences, University of Hyderabad, March 2021.
6. National workshop on Network Science (NetSci2020) held at School of physics, University of Hyderabad, March 2020.
7. 1st TCGA Conference in India on "Multi-Omics Studies in Cancer: Learnings from The Cancer Genome Atlas (TCGA)" held at IISER-Pune, September 2019.

Further, the student has passed the following courses towards fulfilment of coursework requirements for the award of Ph.D.

COURSE NO	TITLE OF THE COURSE	CREDITS	RESULTS
S801	RESEARCH METHODOLOGY/ ANALYTICAL TECHNIQUES	4	PASS
S802	RESEARCH ETHICS, BIOSAFETY, DATA ANALYSIS	4	PASS
S803	RESEARCH PROPOSAL AND SCIENTIFIC WRITING	4	PASS

Supervisor

Dr. VAIBHAV VINDAL
Associate Professor
Dept. of Biotechnology & Bioinformatics
School of Life Sciences
University of Hyderabad
Gachibowli, Hyderabad-500 046.

Co-Supervisor

Dr. P. Manimaran
Professor
School of Physics
University of Hyderabad
Hyderabad-500 046 (TS) India

Head of the Department

HEAD
Dept. of Biotechnology & Bioinformatics
University of Hyderabad
Hyderabad.

8/2/24

Dean of the School

जीव विज्ञान संकाय / School of Life Sciences
हैदराबाद विश्वविद्यालय/University of Hyderabad
हैदराबाद / Hyderabad-500 046.

University of Hyderabad
(A central University established in 1974 by Act of Parliament)

School of Life Sciences
Department of Biotechnology & Bioinformatics
Hyderabad-500046

DECLARATION

I Thippana Mallikarjuna hereby declare that this thesis entitled "Integrative studies to explore key molecular players involved in cervical squamous cell carcinoma" submitted by me under the supervision of Dr. Vaibhav Vindal is a bonafide research work. I also declare that it has not been submitted previously in part or in full to this University or any other University or Institution for the award of any degree or diploma.

Date : 8/2/24

T. Mallikarjuna

Thippana Mallikarjuna

Regd. No. 18LTPH02

Acknowledgments

I want to thank my **parents** for their sacrifices and efforts to give me a good life and a quality education. They have been my pillars of strength and support in difficult times. Moreover, I am indebted to my nanna, who was a great source of inspiration and love for me. I hated losing him on this journey, but he has and will live on in my memories until my last breath. I miss you, **Nanna**.

I express my deepest sense of gratitude to my mentor, **Dr. Vaibhav Vindal**, for giving me an opportunity to work under his able guidance, providing the lab facilities and constant support throughout my work. His continuous encouragement, stimulating guidance, moral support and freedom enabled me to complete research work successfully. The help and advice and constructive criticism given by him time to time shall carry me a long way in the journey of life on which I am about to embark.

I would like to thank my co-supervisor **Prof. P. Manimaran** for his valuable guidance, continuous encouragement and supervision throughout my research.

I would like to thank my Doctoral Research Committee members **Dr. Sunanda Bhattacharyya** and **Prof. Nooruddin Khan** for their cordial support, valuable suggestions and guidance throughout my research work.

I extend my sincere and heartfelt gratitude to Head of the Department of Biotechnology and Bioinformatics **Prof. J.S.S. Prakash** and former Heads, **Prof. K.P.M.S.V. Padmasree**, **Prof. Anand K. Kondapi** and **Prof. Niyaz Ahmed** for providing the necessary departmental facilities for the smooth conduction of research work.

I am thankful to the Dean, **Prof. Anand K. Kondapi**, and former Deans, **Prof. N. Siva Kumar**, **Prof. S. Dayananda**, **Prof. K.V. A. Ramaiah** and **Prof P. Reddanna** School of Life Sciences, University of Hyderabad, for giving me the opportunity and support throughout my work.

I am always thankful to **Prof. Rajagopal Subramanyam** for his constant encouragement and moral support throughout my academic journey at the University of Hyderabad. He always motivated me to pursue excellence in my field of study and helped me overcome many challenges and difficulties.

I would like to express my gratitude to **Prof. P Prakash Babu** and **Prof. P Reddanna** for their valuable suggestions and guidance in all aspects of my research. I appreciate **Prof. P Prakash**

Babu for the opportunity he has provided to collaborate with other researchers in computational projects during my master's thesis at his laboratory. He has been very supportive and encouraging of my talent and potential, and he generously offered me authorship in publications. These experiences helped me to be independent, gain confidence and skills in my field.

I extend my sincere thanks to all the **faculty members** of School of Life Sciences, for permitting me to use the research facilities.

My special thanks to all **non-technical staffs**, for their co-operation throughout my study.

I am further thankful to **Dr. Naresh Damuka**, **Dr. Neelesh Babu Thummadi**, **Dr. Praveen Kumar Simhadri, Dr. Naidu Babu** and **Gouthami Suma** for their assistance in my research work. They have been very supportive and helpful throughout my academic journey.

I gratefully acknowledge all my former and present **members of computational functional genomics laboratory** for their friendly association, encouragement, suggestions and help offered to me. I especially appreciate the camaraderie and joy that **Ayushi**, **Nikhith**, **Aswini**, **Pragya**, **Srija** and **Divya** brought to the lab every day. These guys have been more than labmates.

I acknowledge the **ICMR** for providing me senior research fellowship. Also, **DBT**, **ICMR** and **IoE-UoH** for funding the lab.

I would like to express my sincere gratitude to **DST-SERB**, **DBT-CTEP**, **NUS**, **ISCB** and **Institution of Eminence**-University of Hyderabad (IoE-UoH) for their generous support in the form of travel awards. These awards enabled me to attend and present my research at various international conferences and workshops, which were invaluable opportunities for learning, networking and collaboration. And also for having some fun along the way, because who doesn't like to travel and explore new places? I'm sure my research benefited from the occasional sightseeing and cultural immersion, as well as the exposure to different perspectives and feedback.

I would like to express my sincere thanks to **Dr. Rajagopalan Pandey** and **Rohit Reddy T** for their assistance during my visit to NCKU, Taiwan. They made my trip much easier and enjoyable, as it was my first international academic trip. I learned a lot from them and I appreciate their hospitality and generosity.

I would like to express my gratitude to my friends, especially my bestie **Aruna Varma G**, who has always supported me through thick and thin since our school days. I also appreciate the help and

guidance of **Prasanthi**, **Ravi**, **Abhilash**, **Naveen**, **Kanna**, and **Teja**, who have been with me on this journey. Last but not least, I want to thank my seniors from PG days **Dr. Balaji** and **Dr. Srinu**, for their valuable feedback and advice.

I would like to express my profound appreciation to my father, late **Thirupaiah** garu, mother, **Venkatasubbamma** garu and my siblings, **Mani**, **Sankar**, **Suri** and **Syamala** for their constant encouragement and assistance throughout my journey. I am also immensely grateful for the unconditional and immeasurable love and support from **Prasanna**, **Jessy** and **Sanvi**.

I would like to express my sincere gratitude to my friends who made my stay in HCU a memorable and enriching experience. **Dr. Saleem**, **Kiran**, **Anil**, **Sanjay**, **Subhan**, **Aradhana**, **Netrika**, **Bindia**, **Aswini**, **Shubhangi** and **Pragya** have been supportive, helpful and inspiring throughout my academic journey. They have shared their insights, perspectives and knowledge with me, as well as their joys and sorrows. They have been more than friends; they have been family. I thank them from the bottom of my heart for being there for me.

I am grateful to all the teachers who have taught me throughout my academic journey, from school to postgraduate level (IMSc). I would like to express my special appreciation to **K. Papaiah**, **G. Chakradhar Raju**, **Narayana Raju** and **Dr. C. Madhava Reddy**, who have inspired and motivated me to pursue my passions and goals. They have not only shared their knowledge and skills with me, but also provided me with valuable guidance and support. I owe them a deep debt of gratitude for their influence on my career and life.

All that I cherish today is the grace of my God. I thank the Almighty for answering my prayers, for granting me the strength, wisdom, knowledge and showering his blessings upon me during research work.

… ***Mallikarjuna Thippana***

Dedicated to my parents

Table of Contents

LIST OF FIGURES AND TABLES

Table No.	Title	Page No.

List of Abbreviations

2D-MF-X-DFA	2D multi-fractal detrended cross-correlation
3'-UTR	3'-Untranslated region
5'-UTR	5'-Untranslated region
%	Percent
aa	amino acids
APID	Agile Protein Interactomes DataServer
apl	average path length
BioGRID	Biological General Repository for Interaction Datasets
CCC gene	Candidate cervical cancer (CCC) gene
CCDB	Cervical cancer gene database
CDS	Coding sequence
ceRNA	Competing endogenous RNA
CESC	Cervical Squamous Cell Carcinoma and Endocervical Adenocarcinoma
CGR	Chaos game theory
circRNA	Circular RNA
DAVID	Database for Annotation, Visualization and Integrated Discovery
DEG	Differentially expressed gene
DIP	Database of Interacting Protein
ENCODE	Encyclopedia of DNA Elements
fCGR	frequency CGR
GEO	Gene Expression Omnibus
GLOBOCAN	Global Cancer Observatory
GO	Gene Ontology
GRN	Gene regulatory network
GTEx	Genotype-Tissue Expression
HGNC	Hugo Gene Nomenclature Committee
HMGA	High mobility group A
HOTAIR	HOX Transcript Antisense RNA
HPA	The Human Protein Atlas
HPV	Human papillomavirus
KEGG	Kyoto Encyclopedia of Genes and Genomes
KCC gene	Known cervical cancer gene
Limma	Linear Models for Microarray Data

lncRNA	Long non-coding RNA
MALAT1	metastasis associated lung adenocarcinoma transcript 1
MRE	miRNA response elements
miRNA	microRNA
mRNA	Messenger RNA
NCBI	National Center for Biotechnology Information
NCG	Network of Cancer Genes
ncRNAs	Non-coding RNAs
NGS	Next-generation sequencing
OS	Overall survival
PPI	Protein-Protein Interaction
PPIN	Protein-Protein Interaction Network
PTM	Post-Translational Modification
PVT1	Plasmacytoma variant translocation 1
RNA	Ribonucleic acid
RNA-Seq	RNA sequencing
RMA	Robust Multi-array Average
SNHG1	Small Nucleolar RNA Host Gene 1
SNHG14	Small Nucleolar RNA Host Gene 14
STRING	Search Tool for the Retrieval of Interacting Genes/Proteins
TCGA	The cancer genome atlas
TF	Transcription factor
TOM	Topological overlap matrix
TTD	Therapeutic Target Database
UniProtKB	UniProt Knowledgebase
WGCNA	Weighted gene co-expression network analysis
WHO	World health organization

Chapter 1

Introduction & Review of Literature

1.1 Cancer:

Cancer is a comprehensive term encompassing a range of diseases characterized by abnormal cell growth and proliferation, resulting in the disregard of normal mechanisms governing cell division and differentiation (Hejmadi M. 2019). Unlike healthy cells, which respond to regulatory signals that control their functions and fate, cancer cells acquire autonomy from these signals, evading processes of cell death and senescence. Cancer can arise in any tissue or organ of the body and can invade adjacent structures, leading to local damage and inflammation. Moreover, cancer cells can detach from the primary tumor and disseminate to distant sites through the blood and lymphatic vessels, forming secondary tumors (metastases) that compromise the functions of vital organs and systems. The ability to spread and colonize different parts of the body is a major factor that renders cancer a life-threatening condition.

One of the hallmarks of cancer is the loss of cellular differentiation, the process by which normal cells acquire specialized functions and structures according to their tissue type. Differentiated cells perform specific roles that are essential for maintaining homeostasis and adapting to environmental stimuli. For instance, muscle cells enable various types of movements, such as skeletal muscle contraction, cardiac muscle pumping, and smooth muscle peristalsis. Another example is alveolar epithelial cells, which facilitate gas exchange between the air and blood in the lungs. However, cancer cells originating from these tissues lose their functional abilities and become more immature and unspecialized. This allows them to evade the normal regulatory mechanisms that control cell growth and division. As a result, cancer cells proliferate uncontrollably and invade other tissues, disrupting their normal functions and causing disease.

1.2 Types of cancers:

The classification of cancer types depends on the origin and differentiation of malignant cells. Some of the major categories of cancer are carcinoma, which originates from epithelial cells that line the skin, organs, and glands; lymphoma, which arises from lymphocytes, a type of white blood cell that is part of the immune system; leukemia, that affects the blood-forming cells in the bone marrow; brain and spinal cord tumors, which develop from the cells of the

central nervous system; and sarcoma, which is derived from connective tissue, such as bone, cartilage, muscle, and fat.

Carcinoma is the most prevalent type of cancer in humans, affecting organs such as the lungs, breast, prostate, colon, rectum, and pancreas. Carcinoma cells can either spread to other parts of the body (metastasis) or remain within the original site (in situ). They can be classified into three types based on the degree of invasion: carcinoma in situ, invasive, and metastatic. Carcinoma in situ is a stage of cancer in which abnormal cells are confined to their origin and do not invade other tissues. Invasive carcinoma is a stage in which abnormal cells break through tissue boundaries and invade adjacent tissues. Metastatic carcinoma is a stage in which abnormal cells spread from the primary site to distant tissues/organs through the bloodstream or the lymphatic system.

Depending on their source and location, epithelial cells can develop into various carcinomas. Some examples are basal cell, squamous cell, renal cell, ductal carcinoma in situ (DCIS), and invasive ductal carcinoma (IDC). These cancers affect different epithelial tissues in the skin, organs, and glands. Adenocarcinoma originates from glandular cells that produce mucus or other substances. It can affect many organs, such as the esophagus, lungs, breast, pancreas, prostate, colon and rectum. Adenocarcinoma can grow locally or spread to distant sites.

1.3 Cervical cancer:

Cervical cancer (CC) is a prevalent malignancy that affects women globally. According to the age-adjusted rates from 2015 to 2019, the annual incidence and mortality of CC were 7.8 and 2.2 per 100,000 women, respectively. In 2020, an estimated 604,127 women were diagnosed with CC worldwide. India has a high burden of CC, with approximately 1,23,907 new cases and 77,348 deaths annually. However, these numbers may have been underestimated because of underdiagnosis and underreporting. The age-adjusted incidence rate of CC in India is 18 per 100,000 women, with a cumulative risk of 2.01 percent (Sung et al., 2021).

Moreover, CC is responsible for 17% of all cancer-related mortalities among Indian women aged 30-69 years. Cervical cancer prevalence rates declined by more than half between the mid-1970s and the mid-2000s after the enhanced screening programs, which can detect cervical changes before they become cancerous. In general, incidence rates were stable between 2009

and 2018. Despite the decline in mortality and incidence over the last three decades, CC remains a significant public health challenge in India.

The most important etiological factor for CC is persistent infection with high-risk human papillomavirus (HPV) types, which are responsible for more than 90% of CC cases worldwide. However, HPV infection alone is not sufficient to induce malignant transformation and progression of CC. The interplay of various genetic and epigenetic alterations in both the coding and non-coding regions of the genome influences the development and outcome of CC. Patients with cervical cancer often have poor prognosis due to tumor metastasis and recurrence. Current treatments, such as surgical resection and chemotherapy, are ineffective. Despite advances in understanding the molecular mechanisms of CC, they have not been translated into clinical practice. Thus, it is essential to elucidate the molecular mechanisms involved in CC development and to devise novel therapeutic strategies. One of the molecular mechanisms that has attracted attention is the regulatory role of non-coding RNAs (ncRNAs) in cancer.

1.4 Non-coding RNAs

Non-coding RNAs (ncRNAs) do not encode proteins but play essential roles in regulating various biological processes (Li et al., 2021; Huang et al., 2021 & Fu et al., 2014). Advances in next-generation sequencing (NGS) have enabled researchers to explore the complex transcriptional landscape of tissues and uncover ncRNAs' involvement in carcinogenesis and cancer progression. Cancer affects the expression of both coding and non-coding RNAs in the human genome, which accounts for 2% and 98% of the genome, respectively (Du & Che 2017; Derrien et al., 2012 & Mattick & Rinn 2015). These expression changes influence cellular functions that are regulated by various factors, such as epigenetic regulators, transcription factors, translation factors, and non-coding RNAs (Lee & Young, 2013). Non-coding RNAs such as miRNAs, lncRNAs, small non-coding RNAs, and pseudogenes are essential for regulating gene expression and modulating various cellular processes during development and disease. MicroRNAs (miRNAs) are a well-studied type of ncRNA that have potential applications as biomarkers in cancer (Rasool et al., 2016; Lin & He, 2017). On the other hand, lncRNAs are a relatively novel and distinct class of molecules that have important functions in cancer (Huarte, M. 2015).

LncRNAs can originate from various genomic locations and interact with other genes in different ways. Some lncRNAs are transcribed from enhancers, promoters, or introns of protein-coding genes. In contrast, others are antisense to protein-coding genes and can overlap with them to varying degrees (divergent, terminal, or nested). Additionally, some lncRNAs are derived from pseudogenes or contain one or more hairpins (small RNAs) within their transcripts (Kung et al., 2013). In both the nucleus and the cytosol, lncRNAs have various roles in regulating gene expression. In the nucleus, they modulate epigenetic modifications, transcriptional regulation, splicing events, enhancer activity, protein scaffolding, and chromosomal interactions (Sun et al., 2018; Batista & Chang, 2013). In the cytosol, they affect mRNA stability and translation efficiency (Rashid et al., 2016). Furthermore, lncRNAs can act as miRNA sponges or competing endogenous RNAs (Kallen et al., 2013; Yan et al., 2015; Ma et al., 2014). Moreover, some lncRNAs can be translated into small peptides (Bazzini et al., 2014).

Unlike protein-coding gene transcripts, which have a high degree of sequence conservation and expression across species, lncRNAs are characterized by low sequence conservation and expression levels. This reflects their diverse and context-specific roles in regulating various biological processes (Derrien et al., 2012). Moreover, lncRNAs exhibit strong tissue specificity and influence the transcriptional level alteration of chromatin biology and gene regulation. Although many lncRNAs have been discovered, only a few have been fully characterized. Contrary to the earlier assumption that lncRNAs are non-functional byproducts of transcription due to their low expression, evidence from the past five decades reveals that they play vital functions in regulation of cellular processes including carcinogenesis and metabolism (Ohno 1972). Previous research has indicated that the development and progression of malignant tumors are influenced by long non-coding RNAs (lncRNAs). In cervical cancer, numerous lncRNAs are involved in diagnosing and suppressing metastasis through their expression levels. These lncRNAs have altered levels of expression in relation to the diagnosis and prognosis of treatment response. They can act as either oncogenes or tumor suppressors, making them significant players in cancer studies (Aalijahan & Ghorbian, 2019; Sun et al., 2013).

MicroRNAs (miRNAs) is about 19 to 25 nucleotides long and mostly located in the cytoplasm. They controls gene expression by binding to specific sequences in the 3′ untranslated region (UTR) of mRNA. miRNAs can have both oncogenic and tumor suppressor functions and have been shown to play a crucial role in the initiation and progression of cervical cancer, as well as its metastasis (Wang & Chen 2019). LncRNAs and miRNAs can interact to form complexes, with lncRNAs primarily located in the nucleus and miRNAs in the cytoplasm. These interactions are essential for modulating the expression of oncogenes and tumor suppressors, influencing cancer initiation and progression. LncRNAs and miRNAs regulate gene expression by binding to mRNA. They have a reciprocal relationship, as they share common mRNA targets and their cross-regulation affects gene expression and metastasis. LncRNAs can act as ceRNAs, sequestering miRNAs and preventing them from degrading mRNAs, thereby enhancing translation (positive regulation). In contrast, miRNAs can bind to mRNAs and induce their decay, inhibiting translation (negative regulation) (Berti et al., 2021; Kung et al., 2013).

Our study focused on coding and non-coding RNAs, especially lncRNAs. We integrated the transcription profiles of these two types of RNAs from cervical cancer patient's data to understand their regulatory mechanisms.

1.5 Gene prioritization

Many studies have aimed to identify the genes and pathways responsible for cancer phenotypes. Gene prioritization is a technique that sorts genes based on their relevance and importance for a specific disease or phenotype. It helps to select a subset of genes from an extensive list of candidates that are most probably involved in the disease mechanism and that require further experimental validation. There are various techniques for gene prioritization, such as text mining, machine learning, network-based methods, and hybrid methods (Tranchevent et al., 2011; Kaushal et al., 2020; Azadifar and Ahmadi, 2022). These methods use different types of data and strategies to rank the candidate genes, such as functional similarity of sequences, protein-protein interaction networks, mutational profiles, gene ontology, disease ontology, human phenotype ontology, etc.

Based on the text-mining methods and data sources, the publicly available tools for gene prioritization or identification can be grouped into different categories. Some methods, such as POCUS, PROSPECTR, Gentrepid and PhenoPred, rely on structured data without text mining to prioritize genes. Other methods use text mining to extract relevant information from the literature or databases. These methods can be further classified into four categories: keyword searches, vector space models, ontology structures, and statistical text mining. Keyword search methods like GeneSeeker, Prioritizer, CANDID, PGMapper, GeneProspector, and MaxLink use simple queries to identify genes related to a phenotype or disease. Vector space model methods, such as G2D, SNPs3D, MimMiner, Endeavour, CAESAR, ToppGene, CIPHER, GeneDistiller, PRINCE, PolySearch, GeneWanderer and GPsy, represent genes and phenotypes as vectors in a multidimensional space and compute their similarity or distance. Ontology structure methods, such as Tiffin et al., SUSPECTS and MedSim, use the hierarchical structure of ontologies such as Gene Ontology or Human Phenotype Ontology to measure the semantic similarity between genes and phenotypes. Statistical text mining methods, such as GRAIL, Genie and MetaRanker, depends on advanced techniques such as machine learning, natural language processing or network analysis to identify associations between genes and phenotypes from large-scale text corpora (Luo et al., 2014).

The functional similarity approach leverages the existing knowledge of known cancer genes to discover new genes that have similar sequences or functions. Several methods have been developed to compare protein sequences and to infer their functional similarities. Sequence similarity analysis can be categorized into two types: alignment-free and alignment-based methods. Alignment-free methods compare and analyze DNA or protein sequences without relying on conventional sequence alignment methods. They have several benefits, such as computational efficiency, scalability, and independence from prior knowledge of sequence similarities or homologies. They can also detect homologous regions, functional motifs, and phylogenetic relationships among sequences. Some of the extensively used alignment-free sequence analysis methods are as follows:

k-mer frequency analysis: This technique counts the frequency of all possible k-mers (substrings of length k) in a sequence. The resulting k-mer frequency vectors can then be used to measure the similarity and diversity of sequences.

Composition-based methods use statistical analysis to examine the distribution of specific nucleotides or amino acids within a sequence. They can reveal the compositional features and evolutionary patterns of sequences.

1.6 Biological networks

Graphs or networks are mathematical representations of real-life systems, where the nodes and vertices represent entities and the edges represent the relationships among them. Edges can be either 'directed' or 'undirected,' depending on whether they have a direction. Web networks, scientific collaboration networks, social networks, football game networks, and biomolecular interaction networks are some of the network types that have attracted a lot of research interest. For instance, a scientific collaboration network consists of authors as nodes, and edges that represent co-authorship of a research article between two nodes. In biology, various types of networks involve interactions between different biomolecules, such as protein-protein interaction networks, metabolic networks, gene regulatory networks, signaling networks, etc. Many genes and their regulators interact with each other in a gene regulatory network. A metabolic network has metabolites as nodes and biochemical reactions as edges that transform them. Signaling networks consist of molecules that are linked if they share the same signaling pathway. These networks have several intrinsic properties and applications in solving biological problems.

Network approaches have been used to investigate prognostic value of genes based on their system-level properties in different types of cancer. Protein-protein interaction (PPI) networks capture the interrelated nature of biological processes (Milenkovic et al., 2010). PPI networks can help identify prognostic genes and drug targets in cancer and reveal novel cancer gene mechanisms (Amala & Emerson, 2019; Li et al., 2017). These genes tend to form modules within gene co-expression networks rather than being hub genes. This pattern is specific to each cancer type; however, some modules are conserved across various cancers (Yang et al., 2014). lncRNAs are key regulators of gene regulation in cancer and exhibit a consistent pattern of regulation across different cancer types (Saleembhasha & Mishra 2019). These lncRNAs often correlate with key driver mutations, suggesting their potential roles in cancer progression (Ashouri et al., 2016). They play critical roles in oncogenesis, tumor metastasis, and tumor

suppression (Nie et al., 2012). lncRNAs affect gene expression by interacting with different molecular processes such as transcriptional, post-transcriptional, epigenetic and translational regulation (Sun et al., 2017).

1.7 Network structural topological properties

Various metrics, i.e., centrality measures, are used to define a node's importance based on its topological importance in a network. Some of the frequently used centrality measures are degree, betweenness and closeness. The figure below illustrates a graph G = (V, E), where V is the all vertices and E is the edges that connect them.

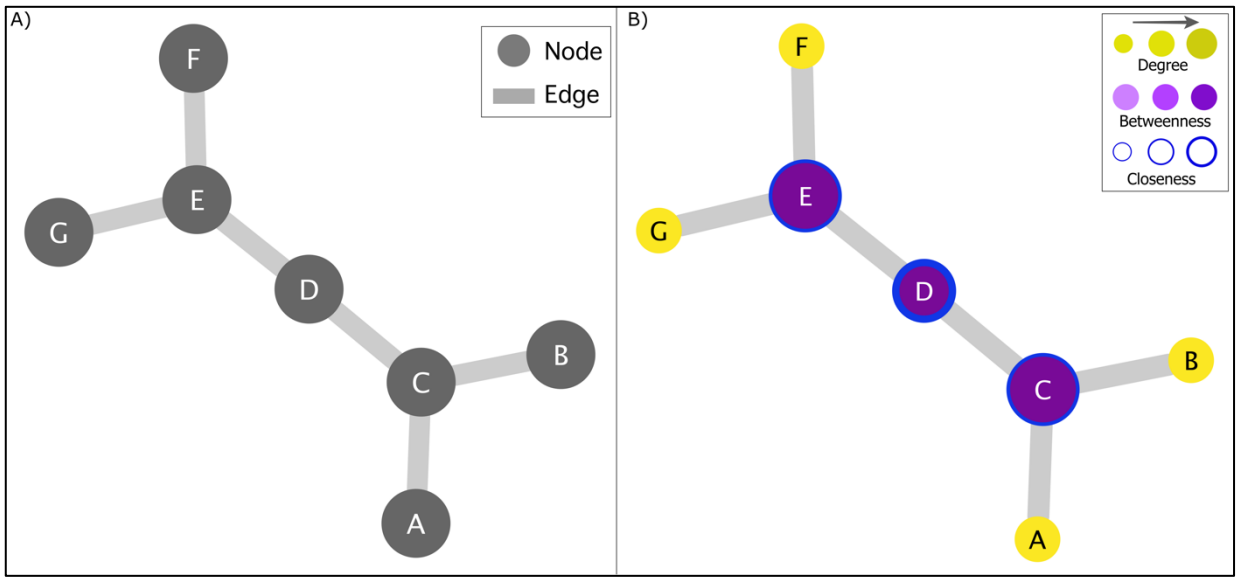


Figure 1: Network/graph with nodes and edges. A) Graph G=(V, E), V has nodes A-G and E has edges between nodes. B) The node color, size, and node border thickness denote the nodes' degree, closeness and betweenness, respectively.

Degree Centrality

A node in a network has a degree as one of its basic attributes, which indicates how many connections the node shares with other nodes in the network. The degree centrality *Cd (v)* of vertex v in graph G = (V, E) can be written as

$$C_d(v) = \frac{\deg(v)}{(N-1)}$$

Where N is the number of vertices in G and *deg(v)* is the number of edges incident to *v*. In network analysis, a node with more connections with other nodes is called a "hub." The number of connections that enter a node is called its "in-degree" and the number of connections that leave a node is called its "out-degree."

All real-world networks exhibit the 80/20 principle, which states that 80% of the results are due to 20% of the effort. In biology, this principle states that a few proteins (20%) perform most of the cell's functions (80%) and regulate most of the cellular processes. It is often applied to identify key nodes, such as hubs and bottlenecks, in biological networks (PPIN & Regulatory networks).

Betweenness Centrality

Betweenness centrality is a metric that shows how important a node is in a network. It counts how many shortest paths between other nodes include that node. A high betweenness centrality means the node connects different parts of the network and affects the information flow.
The betweenness centrality of a node v is given by the expression:

$$C_b(v) = \sum_{\{s \neq v \neq t\}} \frac{\sigma_{\{st\}(v)}}{\sigma_{\{st\}}}$$

Where σ_{st} is the total number of shortest paths from node s to node t and σ_{st} (v) is the number of those paths that pass through v. A node that has the most shortest paths going through it is known as a "bottleneck" node.

Closeness Centrality

Closeness centrality measures how central a node is in a network. It is calculated as inverse of the total distance from that node to every other node. The closeness centrality increases as the node gets closer to all other nodes in the graph. It can be used to analyze the node's efficiency, accessibility, or influence in a network.

The Closeness centrality of a node v can be represented as

$$C_c(v) = \frac{1}{\text{Average distance from } a \text{ node to all other } nodes}$$

Clustering Coefficient

A clustering coefficient quantifies how well the neighbors of a node tend to form clusters or groups. It is two types: global and local. The global clustering coefficient calculates the ratio of closed triplets (three nodes that are all connected) to the overall number of triplets in the network. The local clustering coefficient calculates the fraction of possible connections among the neighbors of a node that are actually present. Both coefficients range from 0 to 1, where 0 means no clustering and 1 means perfect clustering.

It is computed by dividing the number of actual links among the neighbors by the number of potential links that could exist between them. The formula for the clustering coefficient of a node i is:

$$C_i(v) = \frac{2L_i}{k_i * (k_i - 1)}$$

Where L_i is the number of links between the neighbors of node *i* and k_i is the degree of node i (the number of neighbors it has).

Average degree

The network's average degree is the sum of all node degrees divided by the node count. It can be calculated as

$$< C_d > = \frac{1}{N} \sum_d C_{d(v)}$$

Where N represents the overall number of nodes and $C_d\ (v)$ represents the node degree *v* in the network.

Average clustering coefficient

A network's average clustering coefficient is the average value of the clustering coefficients for each node in the network and indicates the network the modularity. It is given as

$$< C_i > = \frac{1}{N} \sum_i C_{i(v)}$$

Where, N represents the overall number of nodes and $C_i\ (v)$ represents the clustering coefficient of the node v in the network.

Average path length

The average path length is computed by finding the shortest distance between every pair of nodes and then taking the mean of those distances.

It is denoted as

$$apl = \sum_{\{x \neq v \neq t\}} \frac{\sigma_{\{st\}(v)}}{N-1}$$

Where N is the overall number of nodes in the network, σ_{st} is the total number of shortest paths from node s to node t and $\sigma_{st}\ (v)$ is the number of those paths that pass through v.

Degree distribution

The degree distribution of a network is the fraction of nodes that have a certain degree 'k', which is the number of connections or edges that a node has (Barabási & Oltvai 2004). The degree distribution reveals the structure and properties of a network, such as its robustness, resilience, and efficiency.

The degree distribution of all real-world networks obeys a power law, meaning that most nodes have a low degree, but there are some nodes with very high degree, called hubs. These hubs are important for the connectivity and robustness of the network, as they link many other nodes together. A power law degree distribution can be written as:

$$P(k) \sim Ck^{-a}$$

Where C is a normalization constant that ensures that the sum of P(k) over all possible values of k is equal to 1.

1.8 Datasets

In this work, multiple datasets were analyzed to prioritize and identify candidate genes associated with cervical cancer. One of the datasets was TCGA-CESC (The Cancer Genome Atlas Cervical Squamous Cell Carcinoma and Endocervical Adenocarcinoma), which contains gene expression quantification data of 303 cervical cancer patients as well as 3 adjacent normal tissues. Another dataset, GTEx (Genotype-Tissue Expression), contains gene expression counts from 19 normal cervical tissues (Ectodermal & Endodermal) from healthy donors. GTEx dataset was used to compare the gene expression patterns of normal and cancerous samples and to identify differentially expressed genes. Additionally, I used NCG6.0 (Network of Cancer Genes), which contains both cancer driver and candidate cancer genes, and CCDB (Cervical Cancer Gene Database), which has cervical cancer genes identified and predicted with experimental and computational techniques.

1.9 Objectives of the present study

Based on the background and available literature, we aimed to prioritize and identify the key molecular players involved in the tumour progression of cervical squamous cell carcinoma. The specific objectives of our study were as follows.

Objective 1: To prioritize candidate genes through the Moment of inertia tensor

Objective 2: To integrate chaos game representation and MF-X-DFA to prioritize candidate genes

Objective 3: To identify key genes and pathways through protein-protein interaction networks

Objective 4: To identify regulatory elements related to cervical cancer progression through integrative networks

Chapter 2

Prioritization of candidate genes using the moment of inertia tensor

2.1 Introduction

Candidate gene prioritization involves ranking genes by their relevance to the biological processes of interest and select the most promising ones for further analysis. With breakthroughs in molecular procedures, sequencing technologies, bioinformatics tools and algorithms, understanding the underlying molecular pathways has improved significantly. This has resulted in the identification of numerous cancer genes and ncRNAs. These molecular entities involved in various cellular processes and pathways (Qin et al., 2019; Burk et al., 2017; Wei et al., 2020; Yang et al., 2021; Yuan et al., 2020 & Hindumathi et al., 2014).

Several methods were used to prioritize candidate genes, including ENDEAVOUR, G2D, SUSPECTS, GFSST, and POCUS (Adie et al., 2006; Turner et al., 2021; Perez-Iratxeta et al., 2005; Zhang et al., 2006 & Aerts et al., 2006). As different methods give a set of genes as candidates and validation of these genes becomes extremely costly, resources can't be devoted to this vast number of candidate genes (Zhang et al., 2020). Unfortunately, the accumulated data on candidate genes for cancer is becoming redundant, as there are a minimal number of bioinformatics tools available for prioritizing the candidate genes for cancers, especially cervical cancer. Therefore, it is imperative to employ a bioinformatic method to prune and prioritize genes for further evaluation.

Various methods have been developed to analyze the sequence similarity between protein sequences to understand the functional similarity of the proteins (Randic et al., 2006; Bai & Wang, 2006; Randic et al., 2008; Li et al., 2008; Wen et al., 2009; Li et al., 2009; Liao et al., 2010; Ghosh & Nandy 2011; He et al., 2012 & Tyanova et al., 2018). However, compared to alignment-based methods, alignment-free methods offer more advantages in terms of computational efficiency and accuracy (Zielezinski et al., 2017). Recently, Piotr Wąż and Bielińska-Wąż developed a technique using the concept of moment of inertia tensor for similarity analysis of DNA sequences (Wąż and Bielińska-Wąż 2013). Later, Hou et al. introduced a method that applies the same idea of tensor to measure the sequence similarity between proteins (Hou et al., 2016). The moment of the inertia tensor describes how an object's mass is distributed relative to all three of its rotational axes. It is an alignment-free method that is fast, efficient, and reliable for comparing sequence similarities.

In this study, we focused on prioritizing candidate cancer genes through sequence similarity analysis between known cervical cancer genes (KCCs) and candidate cervical cancer genes (CCCs) using the concept of the moment of inertia tensor. We further analyzed the GO terms for the prioritized CCCs to highlight their possible roles and provide information on the available drug entries.

2.2 Materials and Methods

2.2.1 Data Collection and pre-processing

The list of genes that cause or are involved in cancer were collected from the Network of Cancer Genes (NCG6.0) database, a manually curated repository on systems-level properties of cancer. The gene list consists of 711 known cancer genes and 1661 candidate cancer genes, based on the approach they were identified in various cancer studies (Repana et al., 2019). In addition, we collected gene list datasets related to Cervical Cancer progression from the cervical cancer gene database (CCDB) (Agarwal et al., 2011). It is a manually curated catalog consisting of 537 genes involved in different stages of cervical carcinogenesis. By mapping the list of cervical cancer genes obtained from CCDB with the list of cancer genes present in NCG6.0, 128 genes were identified as common from both databases. Among these 128 genes, 76 were identified as known cancer genes and 52 as candidate cancer genes. Furthermore, these 128 genes were investigated for their association with cervical cancer in DisGeNET (a database of gene-disease associations) and 82 genes were found to have experimentally validated evidence associated with cervical cancer (Piñero et al., 2019). Therefore, 82 genes are considered as known cervical cancer (KCC) genes and the remaining 46 as candidate cervical cancer (CCC) genes.

We retrieved the protein sequences of 128 cervical cancer genes from the UniProt database using the biomaRt library in R. Furthermore, we curated the sequences using the BioStrings library on the R platform to their canonical sequences in the FASTA format (Durinck et al., 2009 & Pagès et al., 2017).

2.2.2 Construction of protein sequences as a 3D model:

Naturally occurring proteins are composed of a polymer chain of amino acids. Twenty standard amino acids determine protein functions. Recently, Hou et al. visualized protein sequences as

a 3D model using a graphical representation approach (Hou et al., 2016). Their approach used the physicochemical attributes of amino acids, such as hydrophobicity and molecular mass as descriptors. This concept is used in the present study.

Based on their hydrophobicity, the amino acids were split into two groups: hydrophobic amino acids HY= [A, C, F, I, L, M, P, V, W, Y] and hydrophilic amino acids HP = [D, E, G, H, K, N, Q, R, S, T]. Each group was further divided into two groups based on their strengths: strong hydrophilic amino acids SP = [D, K, N, R, S]; weak hydrophilic amino acids WP = [E, G, H, Q, T]; strong hydrophobic amino acids SH = [F, I, L, W, Y]; and weak hydrophobic amino acids = [A, C, M, P, V]. A circle of unit radius was divided into four quadrants, with amino acids distributed along the circumference. The first two quadrants contain hydrophobic amino acids, while the other two quadrants contain hydrophilic amino acids. The sequence of amino acids in each quadrant was based on the alphabetical order of their abbreviated names. Each amino acid was given a coordinate represented as $x_i = \cos(2\pi i/20)$, $y_i = \sin(2\pi i/20)$, where i = 1, 2 . . . 20. The relative residue weight of the respective amino acids was attributed to the z-axis coordinate. The heavier amino acids are given +1, and smaller amino acids are given -1 for z-coordinates, as in **Table 1**.

Table 1: List of amino acids with residue weights and X, Y, and Z-axis coordinates.

	Amino acid	**Symbol**	**Residue wt.**	**X**	**Y**	**Z**
Quadrant 1 Weak Hydrophobic	Alanine	A	71.8	0.9511	0.3090	-1
	Cysteine	C	103.14	0.8090	0.5878	-1
	Methionine	M	131.19	0.5878	0.8090	1
	Proline	P	97.12	0.3090	0.9511	-1
	Valine	V	99.13	0.0000	1.0000	-1
Quadrant 2 Strong Hydrophobic	Phenylalanine	F	147.17	-0.3090	0.9511	1
	Isoleucine	I	113.16	-0.5878	0.8090	-1
	Leucine	L	113.16	-0.8090	0.5878	1
	Tryptophan	W	186.21	-0.9511	0.3090	1
	Tyrosine	Y	163.18	-1.0000	0.0000	1
Quadrant 3 Strong	Aspartic acid	D	115.09	-0.9511	-0.3090	1
	Lysine	K	128.17	-0.8090	-0.5878	1

Hydrophilic	Asparagine	N	114.1	-0.5878	-0.8090	-1
	Arginine	R	156.19	-0.3090	-0.9511	1
	Serine	S	87.08	0.0000	-1.0000	-1
Quadrant 4 Weak Hydrophilic	Glutamic acid	E	129.12	0.3090	-0.9511	1
	Glycine	G	57.05	0.5878	-0.8090	-1
	Histidine	H	137.14	0.8090	-0.5878	1
	Glutamine	Q	128.13	0.9511	-0.3090	1
	Threonine	T	101.11	1.0000	0.0000	-1

Using the coordinates of each amino acid, a 3D model for all the genes (both KCC and CCC genes) was created by applying the moment of inertia tensor concept considering the mass m=1. The amino acids represent 3-D Cartesian coordinates, and the center of mass of the Cartesian coordinate system is defined as $\mu x = \frac{\sum_i m_i x_i}{\sum m_i}, \mu y = \frac{\sum_i m_i y_i}{\sum m_i}, \mu z = \frac{\sum_i m_i z_i}{\sum m_i}$

Where x_i, y_i, and z_i are the m_i coordinates. The tensor of moments of inertia is represented as a matrix.

$$I = \begin{bmatrix} Ixx & -Ixy & -Ixz \\ -Iyx & Iyy & -Iyz \\ -Izx & -Izy & Izz \end{bmatrix}$$

The elements of the matrix are

$$I_{xx} = \sum_i m_i \left(\left(y_i^{\mu}\right)^2 + \left(z_i^{\mu}\right)^2 \right); I_{yy} = \sum_i m_i \left(\left(x_i^{\mu}\right)^2 + \left(z_i^{\mu}\right)^2 \right); I_{zz} = \sum_i m_i \left(\left(x_i^{\mu}\right)^2 + \left(y_i^{\mu}\right)^2 \right);$$

$$I_{xy} = I_{yx} = \sum_i m_i \, x_i^{\mu} y_i^{\mu};$$

$$I_{yz} = I_{zy} = \sum_i m_i \, y_i^{\mu} z_i^{\mu};$$

$$I_{xz} = I_{zx} = \sum_i m_i \, y_i^{\mu} z_i^{\mu}.$$

Where $x_i^{\mu}, y_i^{\mu}, z_i^{\mu}$ denote the coordinates of m_i of a Cartesian coordinate system. The centre of the mass was considered to be the origin. The Eigenvalues of the matrix I are calculated, which

are labeled as $\lambda_1, \lambda_2, \lambda_3$. To define a protein sequence, a vector is given as $\vec{v}^{(p)} = \lambda_1, \lambda_2, \lambda_3$ and the similarity is obtained by the Euclidean distance D between the two protein sequences (P^1, P^2). It is represented as $D(P^1P^2) = \|\vec{v}P^1 - \vec{v}P^2\|_2$. This indicates that the distance and similarity are inversely proportional.

2.3 Results

2.3.1 Categorization of cervical cancer genes

All KCC and CCC genes were grouped based on the experimental observation viz. upregulated/overexpressed, downregulated, post-translational modifications (methylation, mutation, amplification, and polymorphism), and unclassified as reported in the CCDB. Of the 82 KCC genes, 20 were upregulated, 11 were downregulated, 17 were post-translationally modified, and 9 were unclassified. However, some genes fall into multiple categories. Similarly, of the 46 CCC genes, 21 were upregulated, 9 were downregulated, 5 were post-translationally modified, and 9 were unclassified. This categorization is represented by the Venn diagrams in **Figure 1**.

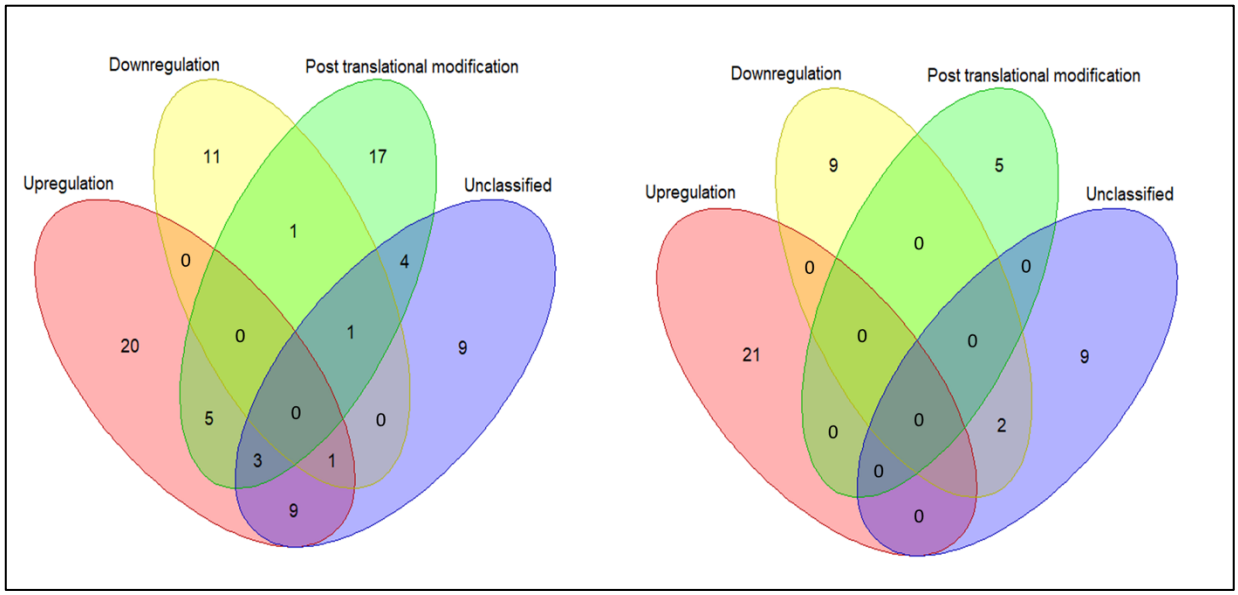


Figure 1: Classification of the proteins based on the experimental observation in known cervical cancer (KCC) genes group and candidate cervical cancer (CCC) genes

2.3.2 Tensor analysis on KCC and CCC proteins

We analyzed the similarity between KCC and CCC proteins using the tensor for the moment of inertia. The largest eigenvalue of the moment of inertia matrix for each protein sequence was considered to calculate the Euclidean distance between any two protein sequences. The resulting distance matrix ranged from 6.133934 to 9834.095, from which we constructed a dendrogram (**Figure 2**).

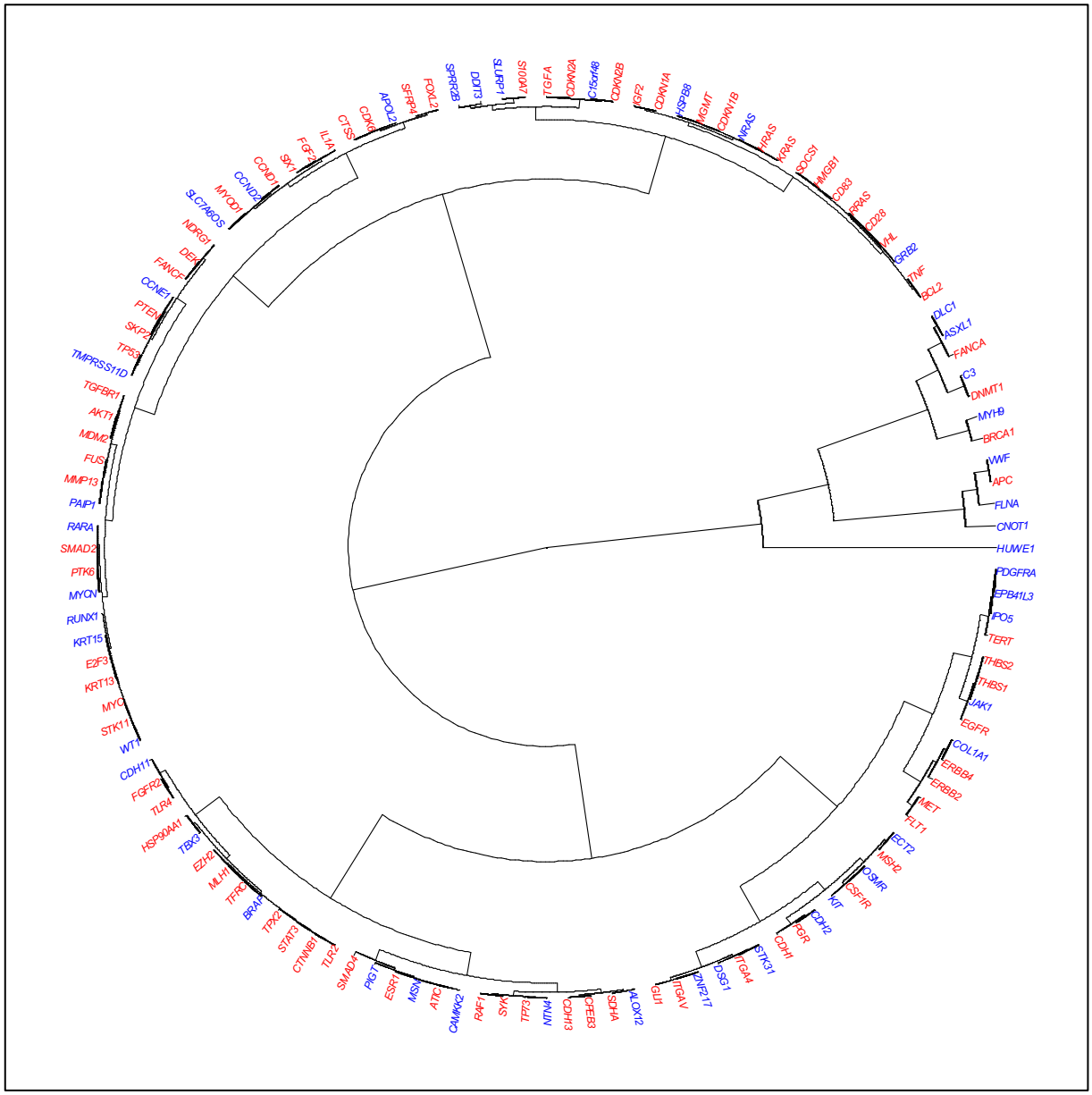


Figure 2: Dendrogram of known and candidate genes belong to cervical cancers

The maximum distance (least similarity) between HUWE1 and S100A7 was observed. Using the distance matrix, we prioritized CCC proteins that showed 1% or less distance (99% similarity or more) from KCC proteins with respect to the maximum distant (least similar) proteins. All CCC proteins except ASXL1, C3, CNOT1, COL1A1, DDIT3, DLC1, FLNA, HUWE1, MYH9, PDGFRA, and VWF showed 99% similarity with one or more KCC proteins with respect to the maximum distance (least similar) proteins.

Furthermore, we considered proteins that showed similarity to more KCC proteins. In this study, we selected proteins that showed at least seven or more associations with KCC proteins (**Table 2**).

Table 2: Prioritized candidate genes for cervical cancer

S.No.	**Gene name**	**Disease**	**Drugs**	**Number of proteins with high similarity**
1	NRAS	Colorectal cancer; Head and neck cancer;	Mutant ras vaccine	**14** (CD28, CD83, CDKN1A, CDKN1B, CDKN2A, HMGB1, HRAS, IGF2, KRAS, MGMT, RRAS, SOCS1, TNF, VHL)
2	GRB2	Acute myeloid leukaemia; Hematologic tumour;	BP-100-1-01	**13** (BCL2, CD28, CD83, CDKN1B, HMGB1, HRAS, IGF2, KRAS, MGMT, RRAS, SOCS1, TNF, VHL)
3	BRAF	Melanoma; Solid tumour/cancer	Dabrafenib	**7** (CTNNB1, EZH2, HSP90AA1, MLH1,STAT3, TFRC, TLR1)

4	CCND2	Not Available	Not Available	**7** (CCND1, CDK6, CTSS, FGF2, IL1A, MYOD1, SIX1)
5	CCNE1	Retinoblastoma;	PD-0183812	**8** (E2F3, KRT13, MYC, NDRG1, PTEN, SKP2, STK11, TP53)
6	RARA	Alzheimer disease; Acute myeloid leukaemia;	Tamibarotene	**11** (AKT1, E2F3, FUS, KRT13, MDM2, MMP13, MYC, PTK6, SKP2, SMAD2, STK11)
7	RUNX1	Not Available	Not Available	**9** (E2F3, KRT13, MYC, PTEN, PTK6, SKP2, SMAD2, STK11, TP53)
8	PAIP1	Not Available	Not Available	**8** (ATK1, E2F3, FUS, MDM2, MMP13, PTK6, SMAD2, TGFBR1)
9	KRT15	Not Available	Not Available	**9** (E2F3, KRT13, MMP13, MYC, PTEN, PTK6, SKP2, SMAD2, STK11)
10	WT1	Acute myeloid leukaemia; Myeloid leukaemia;	WT1-targeted autologous dendritic cell vaccine	**9** (E2F3, KRT13, MMP13, MYC, PTEN, PTK6, SKP2, SMAD2, STK11)
11	MYCN	Not Available	Not Available	**10** (AKT1, E2F3, FUS, KRT13, MDM2, MMP13, MYC, PTK6, SMAD2, STK11)

12	APOL2	Not Available	Not Available	**8** (CCND1, CDK6, CTSS, DEK, FANCF, FOXL2, MYOD1, SFRP4,)
13	HSPB8	Not Available	Not Available	**14** (BCL2, , CD83, CDKN1A, CDKN1B, HMGB1, HRAS, IGF2, KRAS, MGMT, RRAS, SOCS1, TNF, VHL)
14	SLC7A6OS	Not Available	Not Available	**7** (CCND1, CDK6, CTSS, FGF2, IL1A, MYOD1, SIX1)

2.3. 3 Functional enrichment of prioritized proteins

To target a protein, it is imperative to understand its biological processes, molecular functions, and cellular pathways. Functional enrichment analysis was performed using the *clusterProfiler* R package (Yu et al., 2012) for the KCC proteins and prioritized proteins separately to gain insight into the similarity in their functional niche. Statistical significance was set at $p < 0.05$, to enrich the gene ontology terms, such as biological processes, molecular functions, and KEGG pathways.

Biological processes: All KCC proteins showed significance in cell cycles, either positively or negatively regulating several cell types, including epithelial cells, muscle cells, leukocytes, and lymphocytes. However, in CCC proteins, with the p-value adjusted to less than 0.05, only RARA and RUNX1 showed significant biological processes. They both are associated with the regulation of granulocyte differentiation.

Molecular functions: All the KCC proteins showed significant association with molecular functions. The top five molecular functions were phosphatase binding (12), protein phosphatase binding (10), transmembrane receptor protein kinase activity (8), protein tyrosine

kinase activity (9), and protein kinase regulator activity (10). Among the prioritized proteins, 9 showed a significant association with molecular functions. The top five functions included cyclin-dependent protein serine/threonine kinase regulator activity (2), scaffold protein binding (2), translation regulator activity, nucleic acid binding (2), translation regulator activity (2), and insulin receptor substrate binding (1). In addition, the prioritized proteins also have molecular functions, such as protein kinase regulator activity (2), protein kinase B binding (1), neurotrophin receptor binding (1), kinase regulator activity (2), translation activator activity (1), C2H2 zinc finger domain binding (1), translation repressor activity, mRNA regulatory element binding (1), and high-density lipoprotein particle binding (1).

KEGG pathway analysis: With the KEGG analysis, we found that the KCC proteins are mostly involved in various types of cancers. Apart from cancer, some of the KCC proteins are involved in infectious diseases like Human papillomavirus infection (20), Human cytomegalovirus infection (17), Hepatitis B (15), Hepatitis C (14), Epstein-Barr virus infection (15), Kaposi sarcoma-associated herpesvirus infection (14), Measles (12), malaria (6), Salmonella infection (11), Tuberculosis (9), Toxoplasmosis (7), Chagas disease (6), Leishmaniasis (5), Human immunodeficiency virus infection (8), and Coronavirus disease – Covid-19 (6). Among the prioritized proteins, the top five pathways included acute myeloid leukemia (5), chronic myeloid leukemia (4), transcriptional misregulation in cancer (5), prostate cancer (4), and the FoxO signaling pathway (4). In addition, they are involved in various cancers, including gastric cancer (4), endometrial cancer (3), renal cell carcinoma (3), non-small cell lung cancer (3), viral carcinogenesis (4), glioma (3), colorectal cancer (3), breast cancer (3), thyroid cancer (2), hepatocellular carcinoma (3), bladder cancer (2), and melanoma (2). Several proteins are also involved in EGFR tyrosine kinase inhibitor resistance (3), ErbB signaling pathway (3), endocrine resistance (3), neurotrophin signaling pathway (3), PI3K-Akt signaling pathway (4), mTOR signaling pathway (3), chemokine signaling pathway (3), p53 signaling pathway (2), B cell receptor signaling pathway (2), MAPK signaling pathway (3), choline metabolism in cancer (2), cell cycle (2), JAK-STAT signaling pathway (2), and Phospholipase D signaling pathway (2), all of which are involved in cancer development and progression.

2.4 Discussion

Cervical cancer was diagnosed in 5,70,000 women worldwide in 2018, of which 3,11,000 women lost their lives from the disease (Ferlay et al., 2018). Cervical cancer caused by HPV infection usually resolves without symptoms. However, persistent infection can cause cervical cancer, which is lethal in women. Several studies have highlighted the roles of several cervical cancer-causing genes. The genes that are involved in cervical cancer progression are deposited in the DisGeNET database.

The moment of inertia tensor analyzes the sequence similarity as an alignment-free method. Hou et al. (2016) showed the efficiency of this method, where the sequence similarity analysis among the 12 baculoviruses resulted in a similar phylogenetic tree compared to the clustal X-based sequence similarity analysis. Interestingly, in our work, the tensor analysis showed the highest similarity among the KCC proteins that belong to the same family of proteins, such as KRT13 – E2F3; KRAS – HRAS; RRAS – CD28; CD83 - SOCS1; TGFA - CDKN2A; TNF - BCL2. Similarly, we found a high similarity between CCC and KCC proteins that belong to the same family, such as NRAS – HRAS; KRT15 – KRT13; and NRAS – KRAS. This highlights the precision of tensor analysis in analyzing similarities between protein sequences. However, our analysis observed a limitation: this method cannot differentiate any two protein sequences with the same length and amino acid composition but varying in arrangement. However, proteins with similar amino acid compositions are rarely present.

In our study, CCC proteins that are most similar to KCC proteins were prioritized and evaluated further. The final prioritized proteins in correspondence to the GO terms lists NRAS, GRB2, BRAF, CCND2, CCNE1, RUNX1, RARA, KRT15, WT1, MYCN, APOL2, HSPB8, PAIP1, SLC7A6OS.

NRAS belongs to the RAS family, which includes HRAS and KRAS proteins. These proteins are primarily involved in signal transduction. The role of NRAS has been well-established in colorectal cancer, head and neck cancer, acute myeloid leukemia, chronic myeloid leukemia, and melanoma (Wang et al., 2020; Khanna et al., 2015 & Cicenas et al., 2017). Growth factor receptor-bound protein 2 (GRB2) is also involved in signal transduction. It is known to involve prostate cancer, gastric cancer, ovarian cancer, renal cancer, etc. (Ijaz et al., 2017; Qiao et al.,

2020; Ye et al., 2018 & Huang et al., 2018). From the functional enrichment analysis, we found that GRB2 is involved in insulin receptor substrate binding and neurotrophin receptor binding molecular functions. However, for both NRAS and GRB2, there is no significant hit in biological processes. BRAF, a proto-oncogene, is also known as serine/threonine-protein kinase B-Raf. It is involved in cell signaling for cell growth. Several cancers, such as prostate, leukemia, gastric and renal, etc., are well evidenced by the involvement of BRAF (Xue et al., 2018; Steinwald et al., 2020; Vendramini et al., 2019 & Yang et al., 2018). As for molecular functions in GO terms, BRAF has a scaffold protein binding function. CCND2 belongs to the cyclin family of proteins, which are involved in the cell cycle. It is known to be involved in colorectal cancer, ovarian cancer, prostate cancer, etc. (Park et al., 2019; Hua et al., 2019 & Zhu et al., 2014). According to the KEGG pathway analysis, CCND2 is involved in the FoxO signaling pathway, Prolactin signaling pathway, Human papillomavirus infection, PI3K-Akt signaling pathway, cellular senescence, Human T-cell leukemia virus 1 infection, p53 signaling pathway, JAK-STAT signaling pathway. CCND2 molecular functions include cyclin-dependent protein serine/threonine kinase regulator activity and protein kinase regulator activity. CCNE1 is also a cyclin family protein. It is involved in prostate cancer, gastric cancer, etc. (Ju et al., 2019 & Ooi et al., 2017). Similar to CCND1, CCNE1 performs cyclin-dependent protein serine/threonine kinase regulator activity and protein kinase regulator activity. RUNX1 is otherwise called as acute myeloid leukemia 1 protein (AML1). RUNX1 is associated with leukemia and solid tumor growth on the lung, breast, intestine, and skin (Otálora-Otálora et al., 2019). It is involved in the biological process of regulation of granulocyte differentiation. RARA is a nuclear receptor known as NR1B1 (nuclear receptor subfamily 1, group B, member 1). The involvement of RUNX1 in leukemia is well-established (De Braekeleer et al., 2014). It is involved in Th17 cell differentiation and also in the regulation of granulocyte differentiation.

Polyadenylate-binding protein-interacting protein 1 (PAIP1) is involved in translation regulator/activator activity and nucleic acid binding functions. It is associated with breast cancer and cervical cancer (Piao et al., 2018 & Li et al., 2019). KRT15 is a type I cytokeratin involved in the estrogen signaling pathway and performs scaffold protein binding activity. The overexpression of KRT15 is associated with colorectal cancer (Rao et al., 2020). WT1 is a transcription factor with a proline / glutamine-rich DNA-binding domain at the N-terminus and four zinc-finger motifs at the C-terminus. WT1 is inactivated in nephroblastoma and has been

associated with breast cancer (Zhang et al., 2020 & Artibani et al., 2017). MYCN belongs to the MYC family of transcription factors. It is associated with several cancers, such as acute myeloid leukemia, medulloblastoma, and neuroblastoma (Rickman et al., 2018). APOL2 belongs to the apolipoprotein L gene family, which are lipid-binding proteins. It is known to be a biomarker in bladder cancer (Ren et al., 2019). HSPB8 is a heat shock protein, and its role in various cancers is well-studied (Shen et al., 2018 & Crosbie et al., 2013). As all the prioritized proteins are well established in several cancers, these proteins can be studied further for their potential role in cervical cancer and progression. As most of the proteins have drugs available, it would be easier to further explore their efficacy in cervical cancer.

2.5 Conclusion

In this study, the moment of inertia tensor concept was applied to study the sequence similarity between the KCC and CCC proteins and prioritized the potential candidates from the CCC genes list that may play a vital role in cervical cancer progression. With maximum hits in KEGG pathway analysis, the top 5 proteins are NRAS, GRB2, BRAF, CCND2, and CCNE1. As there are drugs available for most of the prioritized proteins, it reduces the efforts in designing the drugs and therapeutic regimes. We found this approach relatively fast and efficient in calculating the similarity between protein sequences. Our work sheds light on the importance of the moment of inertia tensor in prioritizing genes based on sequence similarity. This approach may find applications in sequence similarity analysis in other complex systems.

Chapter 3

Prioritization of candidate genes using chaos game and fractal-based time series approach

3.1. Introduction

Cancer research has increased significantly, but it requires careful analysis to identify suitable markers for diagnosis and treatment. There has been a huge surge of studies on cervical cancer that need to be carefully analyzed for the identification of a suitable marker for diagnostic and therapeutic strategies. The field of bioinformatics has undergone significant progress in recent years. Several studies focused on identifying the candidate genes in cervical cancer using bioinformatics tools. However, we cannot continue to test the efficacy of all the candidate genes in the progression of cancer as it is financially burdening and cumbersome. Therefore, we need to prioritize the candidate genes to test their efficacy.

The ontology-based approach, computation-based approach, and integrated identification approach are generally utilized for candidate gene identification. The biological function of the gene is the basis for gene ontology-based methods. Machine learning, Hidden Markov analysis, data mining analysis, cluster analysis, and KNN classification algorithm are some of the computational methods regularly used to identify candidate genes. Integrative approaches utilize experimental and theoretical data from different sources, such as protein-protein interactions and pathway analysis (Zhu & Zhao 2007). The integration-based methods are generally based on sequence similarity, protein-protein interactions and gene ontology, etc. For example, SUSPECTS, PROSPECTR, and Endeavour prioritize genes by sequence similarity and their function. Sequence similarity holds the key to identify the candidate genes as the sequence of the gene determines the protein sequence, in turn, its function. Owing to the importance of sequence similarity in identifying the candidate genes, in this study, we prioritized candidate genes of cervical cancer using the sequence similarity based on the integrated approach of Chaos theory (CGR) and 2D multi-fractal detrended cross-correlation (2D-MF-X-DFA) and gene ontology.

Chaos Game Representation, an iterating mapping technique, was introduced by Jeffrey (Jeffrey 1990) to represent the genomic sequence in a 2-dimensional image. Several approaches were introduced to visualize the DNA sequences in a graphical way. CGR works on the principle to map a one-dimension sequence to two dimensions or higher space. Jeffrey used a square with four vertices being Adenine (A), Thymine (T), Guanine (G), and Cytosine (C). The algorithm works by drawing a point (P1) at half the distance from the first nucleotide of the

given DNA sequence. Now, considering P1 as the new beginning point, another point (P2) will be drawn at the half distance from the second nucleotide and this process keeps repeating till the end of the sequence. It is noted that DNA sequences show fractal arrangement.

On the other hand, the random numbers did not show any fractal arrangement. Since then, CGR has been favored for alignment-free comparisons of DNA sequences, protein sequences, and phylogeny (Löchel & Heider (2021). In the present study, we represented the image in a square matrix of length L with vertices A(0,0), G(1,1), C(0,1) and T(1,0). The image is represented as dots and spaces, where the dot indicates the position of the nucleotide and the space indicates the absence of the nucleotide. Several ways are present to represent the image, i.e., by changing the order of the vertices. The image also changes with the selected order of vertices. The graphical representations of the genomic sequences have been providing novel insights in deciphering the complexity of genomic sequences as proven in previous studies that include understanding the genome sequences, DNA sequences, RNA sequences, protein sequences, protein-protein interactions, and protein sequence evolution (Tanchotsrinon et al., 2015; Wu et al., 2010; Zhou 2011; Chou 2010; Yu et al., 2004; Deschavanne et al., 1999; Xiao et al., 2010; Dutta & Das 1992 and Lu et al., 2011).

Multifractal nature is seen in nature and also in several social and financial fluctuations (Mandelbrot 1983). Many approaches and techniques were developed to understand the correlation behavior and multifractal nature. Some of the methods are wavelet transform module maxima (WTMM), multifractal detrended fluctuation and discrete wavelet-based fluctuation analysis, etc. (Arneodo et al., 1988; Kantelhardt et al., 2002; Manimaran et al., 2005 and Sahoo et al., 2020). Later, to capture the power-law cross-correlations among two non-stationary datasets with multifractal features, MF-X-DFA was proposed as a method to quantify the multifractal properties of such cross-correlations (Podobnik et al., 2008; Zhou, 2008; Podobnik et al., 2009 and Jiang et al., 2011). The potential of this method has been seen in various events like finance, multifractal random walks (MRWs), climatic changes, seismic events, agricultural future markets, stock market fluctuations, electricity and carbon markets (Pal & Manimaran 2019; Rafique et al., 2022).

In our earlier studies, the coding and non-coding DNA sequences (Pal et al., 2015) and genome sequences (Pal et al., 2016) were analyzed using the integrated approach of CGR and

multifractal detrended analysis. Recently, the mitochondrial genome sequences were also analyzed (Thummadi et al., 2021). In this study, candidate genes of cervical cancer were prioritized using the integrative analysis of CGR and two-dimensional multifractal cross-correlation.

3.2. Materials and Methods

3.2.1 Data collection and pre-processing

The Network of Cancer Genes (NCG6.0) database (Repana et al., 2018) was used to obtain 711 known and 1661 candidate cancer genes. 537 genes involved in cervical cancer were obtained from the Cervical Cancer Gene Database (CCDB) (Agarwal et al., 2011). A comparison showed that 128 genes from CCDB are common with the NCG6.0 gene list. These 128 genes were then analyzed for gene-disease association in DisGeNET (Piñero et al., 2020). It was found that 82 genes out of these 128 genes have experimental validation. Therefore, the experimentally validated 82 genes were considered as known cervical cancer (KCC) genes and the experimentally unvalidated 46 genes as cervical cancer candidate (CCC) genes. The biomaRt R library was used to retrieve gene sequences for the 128 genes from the Ensembl database (Durinck et al., 2009).

3.2.2 Chaos game representation of cervical cancer gene sequences

The fractality of gene sequences was analyzed using chaos game representation (Jeffrey. 1990). The methodology for using the CGR algorithm has been given in detail elsewhere (Jiang et al., 2011; Rafique et al., 2022 and Pal et al., 2015). In brief, the nucleotides A(0,0), G(1,1), C(0,1) and T(1,0) are taken as the vertices of a unit square with length – L. Positions of nucleotides were calculated using the iterative mapping function as given below:

$$P_i = 0.5\ (P_{i-1} + V_{ip}) \quad (1)$$

$$Q_i = 0.5\ (Q_{i-1} + W_{iq}) \quad (2)$$

Where P_i and Q_i are the i^{th} nucleotide co-ordinates computed from half of the previous nucleotide position. The first nucleotide P_{i-1} and Q_{i-1} positions are given from the centre of the unit square (0.5, 0.5). V_{ip} and W_{iq} denote the vertex coordinates. With the iteration of these steps, we have calculated the coordinates for all the nucleotides from all gene sequences and

developed a CGR image. For further analysis, the data matrix was obtained by converting these images to frequency CGR (fCGR).

3.2.3 Two-Dimensional MF-X-DFA technique

The two-dimensional MF-X-DFA method was applied to the fCGR matrix to examine the cross-correlation patterns and multi-fractal properties of gene sequence matrices. The advantage of this integrative approach is that it can accommodate unequal gene lengths for cross-correlation. The 2D MF-X-DFA approach was introduced by W. X. Zhou (Zhou 2008) and the detailed procedure is as follows:

Step 1: Consider any pair of equal-sized two-dimensional data matrices of images i(p,q) and j(p,q), where p=1,2, . . ., a and q=1,2, . . . ,b.

Step 2: The data was split into a_s x b_s non-overlapping square fragments of equal size. For instance, s x s, with a_s=a/s and b_s=b/s. Each data fragment is represented by $i_{x,z}$ or $j_{x,z}$ such that $i_{x,z}(p,q) = i(h_x+p , h_z+q)$ and $j_{x,z}(p,q)=j(h_x+p , h_z+ q)$ for q≤p, q≤s, where h_x =(x−1)s and h_z=(z−1)s.

Step 3: $i_{x,z}$ or $j_{x,z}$ is defined as follows:
$I(p,q) = \sum_{t_1=1}^{p} \sum_{t_2=1}^{q} i_{x,z}(t_1,t_2)$ and $I_{x,z}(p,q) = \sum_{t_1=1}^{p} \sum_{t_2=1}^{q} j_{x,z}(t_1,t_2)$ here q≤p, l≤s.

Step 4: Any set of two fragment's detrended covariance is calculated as:

$$F_{x,z}(s) = \frac{1}{s^2} \sum_{p=1}^{s} \sum_{q=1}^{s} \left[I_{x,z}(p,q) - \tilde{I}_{x,z}(p,q) \right] \left[J_{x,z}(p,q) - \tilde{J}_{x,z}(p,q) \right] (3)$$

Here, $\tilde{I}_{x,z}$ and $\tilde{J}_{x,z}$ represent the polynomial approximations of $I_{x,z}$ and $I_{x,z}$ respectively. The polynomial function was chosen as the least complex plane $\tilde{y}(p,q) = mp + nq + r$, which is used in our analysis.

Step 5: The detrended covariance was used to calculate the qth order fluctuation function $F_{ij}(n,s)$ as shown in step 4 with a square and the mean of all the segments,

$$F_{ij}(n,s) = \left(\frac{1}{a_s b_s} \sum_{x=1}^{a_s} \sum_{z=1}^{b_s} \left[F_{x,z}(s) \right]^{n/2} \right)^{1/n} \quad (4)$$

In accordance with L'Hospital's rule for n=0, the fluctuation function is defined as

$$F_{ij}(n,s) = exp\left(\frac{1}{2a_s b_s} \sum_{x=1}^{a_s} \sum_{z=1}^{b_s} ln\left[F_{x,z}(s) \right] \right) (5)$$

Here, "n" is referred to as an order of the moment and can have any real value.

Step 6: We repeated method steps 2 through 5 using various scale values 's' for varied estimations of 'n'. With the fluctuation function investigation, we obtained power-law scaling behavior.

$$F_{ij}(n,s) \sim s^{h_{ij}(n)} (6)$$

If the calculated scaling examples $h_{ij}(n)$ values do not show a dependence on q esteems, they are of a monofractal nature. If $h_{ij}(n)$ values show dependency on n esteems, then it represents a multifractal nature. However, if i=j, then 2D MF-X-DFA is same as 2D-MFDFA. For the positive 'n' values, $h_{ij}(n)$ indicates large fluctuations, while the negative 'n' values represent small fluctuations.

The strength of the multifractal behavior of cross-correlated image data was analyzed by evaluating the $f_{ij}(\alpha)$ spectrum. $F_{ij}(\alpha)$ values were obtained from Legendre transform $\tau_{ij}(n)$ as:

$$F_{ij}(\alpha) \equiv n\alpha_{ij} - \tau_{ij}(n) \quad (7)$$

Here $\tau_{ij}(n) = nh_{ij}(n) - D_f$.

In the present study, we applied the D_f value as 2. The α_{ij} values were obtained from $\alpha_{ij} = d\tau_{ij}(n)/dn$. The width of the $f_{ij}(\alpha)$ range determines the strength of multifractal behavior. Strong multifractal behavior is indicated by a broader range, while a narrow range indicates weak multifractal behavior.

3.2.4 Functional enrichment and survival analysis

The functional enrichment analysis was performed on candidate genes to get to know about their associated GO terms and KEGG pathways using the clusterProfiler R package (Yu et al., 2012). Also, the candidate genes prognosis was studied via survival analysis using the KM plotter online tool (Lánczky and Győrffy 2021). Survival analysis results show the effect of gene expression pattern levels on patient survival.

3.3 Results and Discussion

Pre-processed nucleotide sequences of known and candidate cervical cancer genes were used to generate the frequency of CGR matrices [Figure 1]. The frequency CGR matrices for each gene were extracted after applying CGR on all the genes. Each CGR image is divided into $2^k * 2^k$ grids, and '*k'* is known to be the length of the DNA segment in the sequence. The CGR analysis was performed by taking *k* as 6, 7, 8, 9, and 10. The results are consistent across different k values and do not show any significant variation, i.e., 6,7,8,9, and with a grid size of 64 x 64.

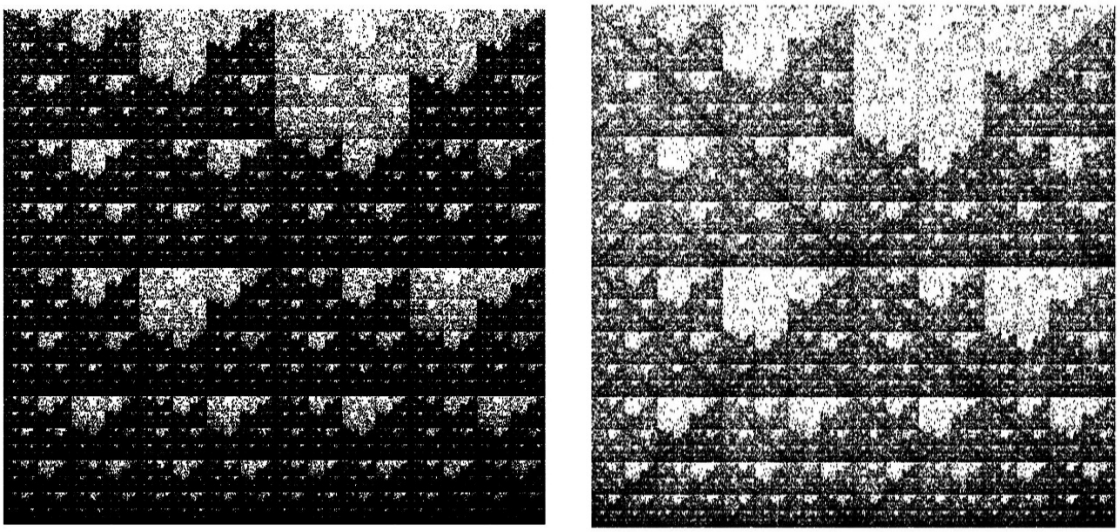


Figure 1: CGR images of ERBB4 (left panel) and BRAF (Right panel)

The fractal behavior and potential candidate genes for cervical cancer were characterized and predicted by proceeding with the k = 6 frequency CGR matrix. It should be noted that the CGR method could generate frequency CGR matrices of equal size even though two sequences are of different lengths.

Further, the 2D MF-X-DFA method was used to characterize multifractal behavior and cross-correlation among the sequences. The cross-correlation was measured among KCC-KCC, KCC-CCC and CCC-CCC sequences. Multifractal nature with varied strength was shown by all the genes [Figure 2]. A scale range (Pal and Manimaran 2019) and also the q^{th} order moments from -10 to +10 value with a step size of 0.2 were used for this study. The results show that the strength of multifractality varies among the sequences depending on the q values. The singularity spectrum's width reflects the multifractality's intensity: a wider spectrum implies a more multifractal behavior, while a narrower spectrum implies a less multifractal behavior [Figure 3]. Additionally, a cluster analysis was performed to find the class affiliation among the cervical cancer genes (known and candidate). These clusters were visualized as circular dendrogram [Figure 4].

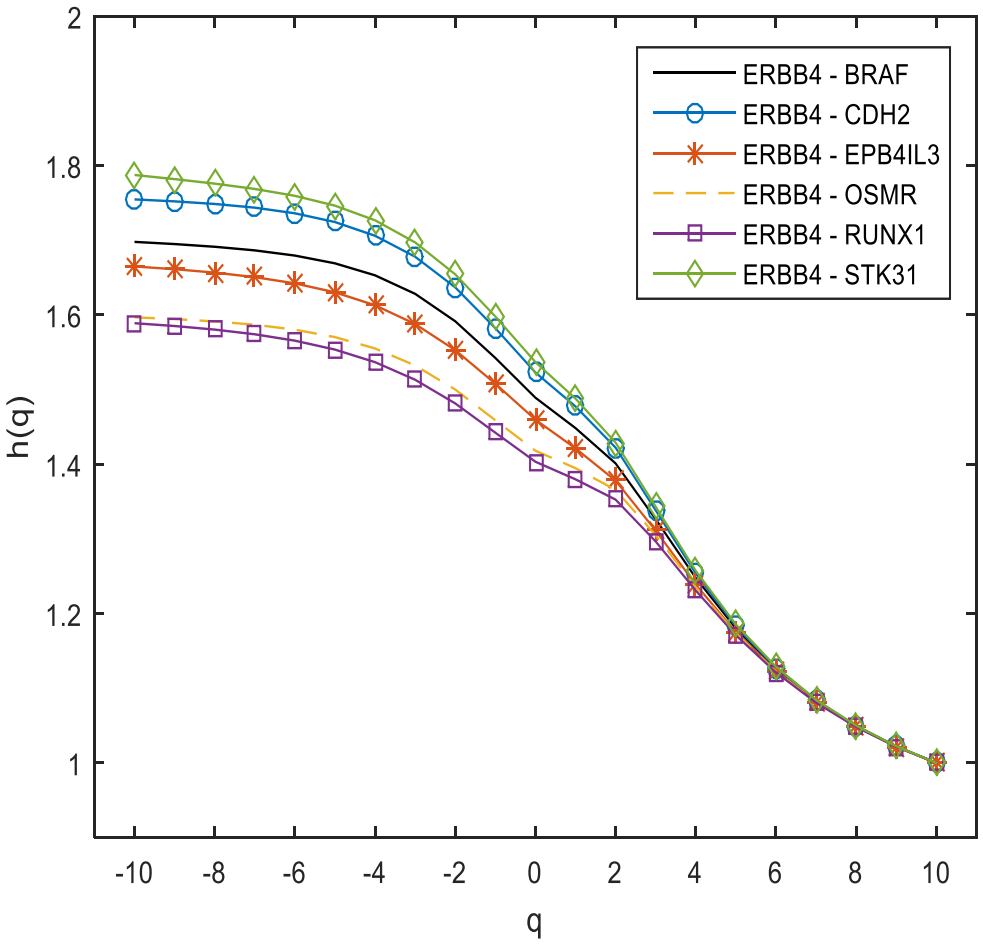


Figure 2: Representative multifractal behaviour of Known cervical cancer gene-ERBB4 and Candidate cervical cancer genes-BRAF, CDH2, EPB4IL3, OSMR, RUNX1, STK31. The values of the h(q) exponents vary according to the choice of q.

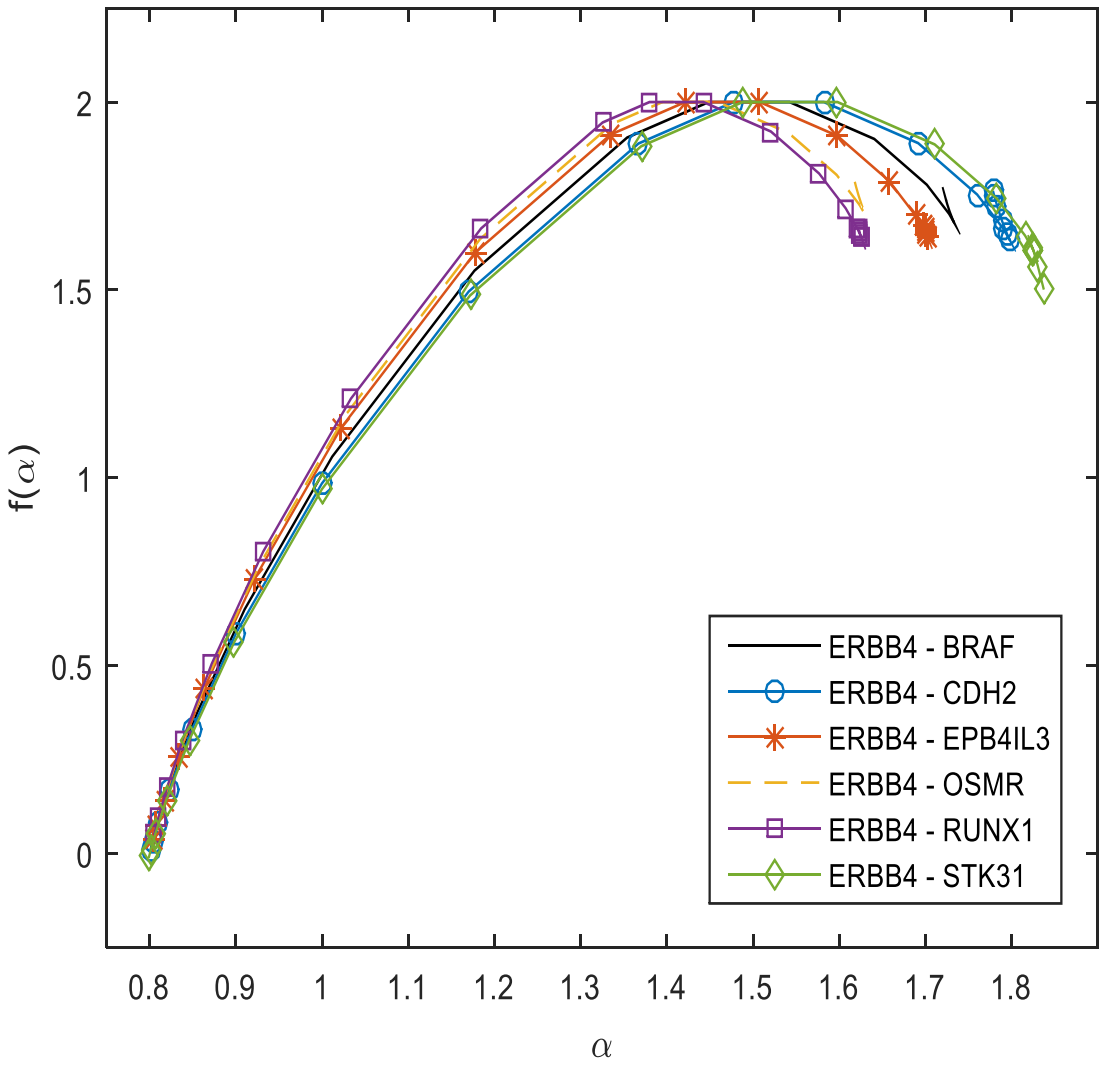


Figure 3: The singularity spectrum f(α) of candidate cervical cancer genes-BRAF, CDH2, EPB4IL3, OSMR, RUNX1, and STK31 in comparison to known cancer gene- ERBB4 showing the strength of the multifractal nature.

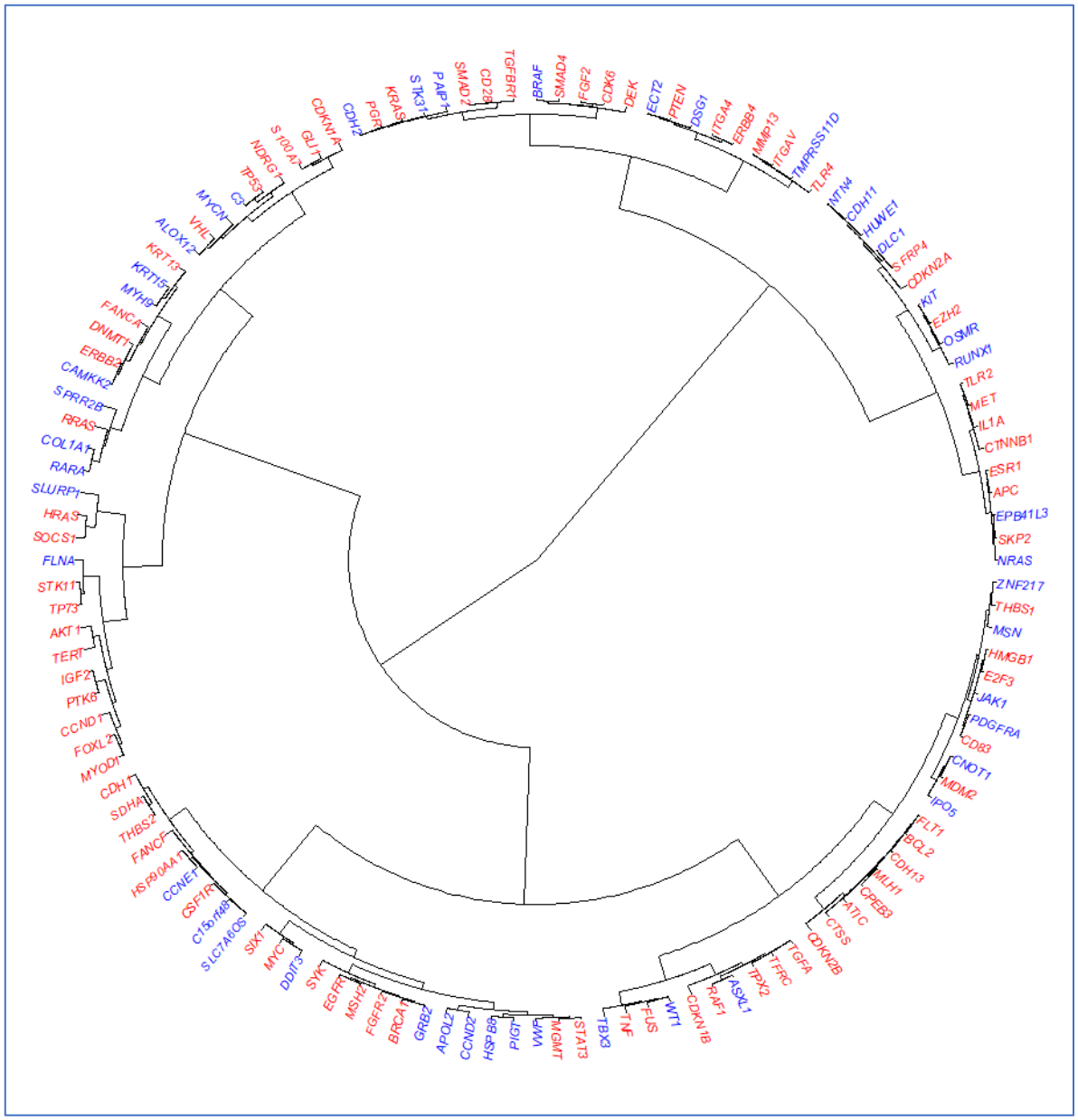


Figure 4: Dendrogram of known and candidate genes belong to cervical cancers

An advantage of alignment-free methods is that a data matrix of the same size can be obtained even when the sequences have different lengths. A total of 16 genes that may be associated with the development of cervical cancer [Table 1] were selected by applying CGR and the 2D MF-X-DFA approaches. A high number of correlations with known cervical cancer genes were found for these candidate genes. The 16 candidate genes are DSG1, ECT2, TMPRSS11D,

STK31, CDH2, PAIP1, BRAF, CDH11, HUWE1, NRAS, DLC1, NTN4, EPB41L3, OSMR, KIT, RUNX1.

Table 1 : Prioritized candidate genes for cervical cancer

S. No.	Gene	Disease	Drugs	No. of Genes	Genes showing similarity
1	DSG1	Not available	Not available	25	APC, CD28, CDK6, CDKN2A, CTNNB1, DEK, ERBB4, ESR1, E2H2, FGF2, IL1A, ITGA4, ITGAV, KRAS, MET, MMP13, PGR, PTEN, SFRP4, SKP2, SMAD2, SMAD4, TGFBR1, TLR2, TLR4
2	ECT2	Not available	Not available	25	APC, CD28, CDK6, CDKN2A, CTNNB1, DEK, ERBB4, ESR1, E2H2, FGF2, IL1A, ITGA4, ITGAV, KRAS, MET, MMP13, PGR, PTEN, SFRP4, SKP2, SMAD2, SMAD4, TGFBR1, TLR2, TLR4
3	TMPRSS11D	Not available	CHEMBL2086421 (Inhibitor 1[Colombo et al., 2012])	24	APC, CD28, CDK6, CDKN2A, CTNNB1, DEK, ERBB4, ESR1, FGF2, IL1A, ITGA4, ITGAV, KRAS, MET, MMP13, PGR, PTEN, SFRP4, SKP2, SMAD2, SMAD4, TGFBR1, TLR2, TLR4
4	STK31	Not available	Not available	21	APC, CD28, CDK6, CTNNB1, DEK, ERBB4, ESR1, FGF2, IL1A, ITGA4, ITGAV, KRAS, MET, MMP13, PGR, PTEN, SMAD2, SMAD4, TGFBR1, TLR2, TLR4
5	CDH2	Solid tumour/cancer	Exherin	19	CD28, CDK6, CTNNB1, DEK, ERBB4, FGF2, IL1A, ITGA4, ITGAV, KRAS, MET, MMP13, PGR, PTEN, SMAD2, SMAD4, TGFBR1, TR2, TLR4
6	PAIP1	Not available	Not available	18	CD28, CDK6, CTNNB1, DEK, ERBB4, FGF2, ITGA4, ITGAV, KRAS, MET, MMP13, PGR, PTEN, SMAD2, SMAD4, TGFBR1, TLR2, TLR4
7	BRAF	Melanoma; solid tumor/ cancer	Dabrafenib	13	CD28, CDK6, ERBB4, FGF2, ITGA4, ITGAV, KRAS, MMP13, PGR, PTEN, SMAD2, TGFBR1, TLR4
8	CDH11	Rheumatoid arthritis	RG6125	6	ERBB4, ITGA4, ITGAV, MMP13, PTEN, TLR4
9	HUWE1	Not available	Not available	6	ERBB4, ITGA4, ITGAV, MMP13, PTEN, TLR4

10	NRAS	Colorectal cancer; head and neck cancer	Mutant ras vaccine	6	ERBB4, ITGA4, ITGAV, MMP13, PTEN, TLR4
11	DLC1	Not available	Not available	5	ITGA4, ITGAV, MMP13, PTEN, TLR4
12	NTN4	Not available	Not available	5	ERBB4, ITGA4, ITGAV, PTEN, TLR4
13	EPB41L3	Not available	Not available	4	ERBB4, ITGA4, ITGAV, TLR4
14	OSMR	Not available	Not available	4	ERBB4, ITGA4, PTEN, TLR4
15	KIT	Tenosynovial giant cell tumour, Metastatic colorectal cancer	Ripretinib	3	ERBB4, ITGA4, PTEN
16	RUNX1	Not available		1	ERBB4

DSG1 (Desmoglein 1) is a cadherin-like transmembrane glycoprotein that is the main component of the desmosome along with armadillo proteins and plakin proteins. Reduction of desmosomal component results in tumor development. Its downregulation is associated with various types of cancers, including those affecting the head and neck, the colon, the skin, the esophagus, the lung, the cervix, and the stomach (Liu et al., 2021). ECT2 (Epithelial Cell Transforming 2) is a guanine nucleotide exchange factor (GEF), which has an essential role in activating Rho family GTPases, thus regulating various cellular processes like cytokinesis, cell division, etc. ECT2 dysregulation is associated with various types of cancer, including those affecting the breast, the lung, and the stomach (Miki et al., 1993 and Chen et al., 2020). TMPRSS11D (Transmembrane Serine Protease 11D) is also denoted as human airway trypsin-like protease (HAT) and is associated with the family of type II transmembrane serine proteases (TTSP). It involves in the proteolytic activation of influenza A, influenza B, and SARS-CoV. Its role is well-established in squamous cell carcinogenesis (Cao et al., 2017). STK31 (Serine/Threonine Kinase 31) is a member of the Serine/Threonine Kinases family. Recent studies reported STK31 as a novel cancer/testis antigen (CTA), CTAs are tumor antigens, ideal targets for cancer immunotherapy (Yokoe et al., 2008). Its role is associated with colorectal and gastric cancers. It is involved in regulating cell cycle progression (Kuo et al., 2014).

CDH2 (Cadherin 2), also referred to as N-cadherin, is the essential factor in the transition of tumors to malignant. It may function as a potential therapeutic target for different cancers (Warde, 2011 and Guvakova et al., 2020). CDH11 (Cadherin 11) is a tumor suppressor gene associated with various tumors. It modulates the activity of AKT/Rho A and Wnt/β-catenin pathways (Li et al., 2012). CDH2 and CDH11 belong to the cadherin family of cell-cell adhesion molecules involved in critical biological processes interacting with each other. They are well known for regulating various characteristics of cell behavior such as differentiation, proliferation, cell polarity, self-renewal, apoptosis, and embryonic stem cell differentiation, and maintenance of tissue integrity (Chen et al., 2021). PAIP1 (Poly(A) Binding Protein Interacting Protein 1) is associated with functions such as translation regulator/activator activity and nucleic acid-binding. It is associated with breast cancer and cervical cancer (Piao et al., 2018 and Li et al., 2019). BRAF (B-Raf Proto-Oncogene, Serine/Threonine Kinase) is involved in cell signaling and it is well established in multiple cancers, such as leukemia, prostate, renal and gastric and so forth,. (Xue et al., 2018; Steinwald et al., 2020; Vendramini et al., 2019 and Yang et al., 2018). HUWE1, also known as E3 ubiquitin ligase, plays a vital role in ubiquitination and proteolysis of target genes. The ubiquitin system dysregulation often lead to pathogenesis, including development of tumors (Kao et al., 2018). NRAS (NRAS Proto-Oncogene, GTPase) is part of the RAS GTPase family comprising HRAS and KRAS. They play an important role in signal transduction pathways. NRAS is well-studied in multiple cancers such as head and neck, acute and chronic myeloid leukemia, colorectal, melanoma, etc. (Wang et al., 2020; Khanna et al., 2015 and Cicenas et al., 2017).

DLC1 (Deleted in Liver Cancer 1) is a Rho GTPase Activating Protein, that acts as a tumor suppressor gene in multiple cancers including lung, prostate, breast and colorectal cancers (Sanchez-Solana et al., 2021). Netrin 4 (NTN4) is part of the neurite guidance factors family associated with neurite growth promotion and elongation. It is a prognostic factor in breast cancer progression (Yi et al., 2022 and Hao et al., 2020). EPB41L3 (Erythrocyte Membrane Protein Band 4.1 Like 3) is a tumor suppressor gene involves in the modulation of the activity of protein arginine N-methyltransferases. It is also, associated with cytoskeleton organization. It is established as a biomarker for meningioma (Zeng et al., 2018). OSMR (Oncostatin M Receptor) is associated with the IL-6 cytokine family and acts as the master regulator in the crosstalk between immune and nonimmune cells. It is a key factor for tumor progression in

breast and ovarian cancers (Araujo et al., 2022). KIT (proto-oncogene c-KIT) is a receptor tyrosine kinase, involved in the activation of MAPK, JAK/STAT and PI3K pathways. It plays a critical role in melanogenesis, gametogenesis, and hematopoiesis (Ke et al., 2016). RUNX1 also known as acute myeloid leukemia 1 protein (AML1). It is involved in leukemia and solid tumor growth in the lung, skin, breast, and intestine (Otálora-Otálora et al., 2019). It plays an essential role in Th17 cell differentiation and regulating granulocyte differentiation. It is well-established in leukemia (De Braekeleer et al., 2014).

Gene ontology and KEGG pathway investigation revealed that the genes under study were related to several biological processes (BP) involving cell-cell junctions, cell shape, and actomyosin structures. Moreover, the genes were involved in various signaling pathways and cancer types, such as Rap1, ErbB, MAPK, PI3K-Akt, mTOR, acute and chronic myeloid leukemia, breast, thyroid, bladder, and gastric cancer. Further, we performed survival analysis on candidate genes using a KM plotter to predict the prognosis in cervical cancer patients. The KM survival curve is widely used statistically to estimate the time to death-events. KM survival curve results were analyzed based on hazard ratio and log-rank p values to filter statistically significant genes that are poorly prognosed in cervical cancer patients. It is found that a total of six genes CDH2, PAIP1, BRAF, EPB41L3, OSMR and RUNX1 have poor prognostic power in patients [Figure: 5].

The candidate cervical cancer genes from our results were prioritize based on sequence similarity with established genes. This provides an advantage to experimental researchers in designing and developing new drugs and antibodies against multiple target molecules as a comprehensive approach. Moreover, alignment-free methods have an edge over alignment-based methods for sequence similarity. However, our goal is limited to prioritizing candidate cancer genes based on sequence similarity. Furthermore, this approach may find applications in predicting new cancer genes, differentiating driver and non-driver cancer genes, clustering and classification problems, etc.

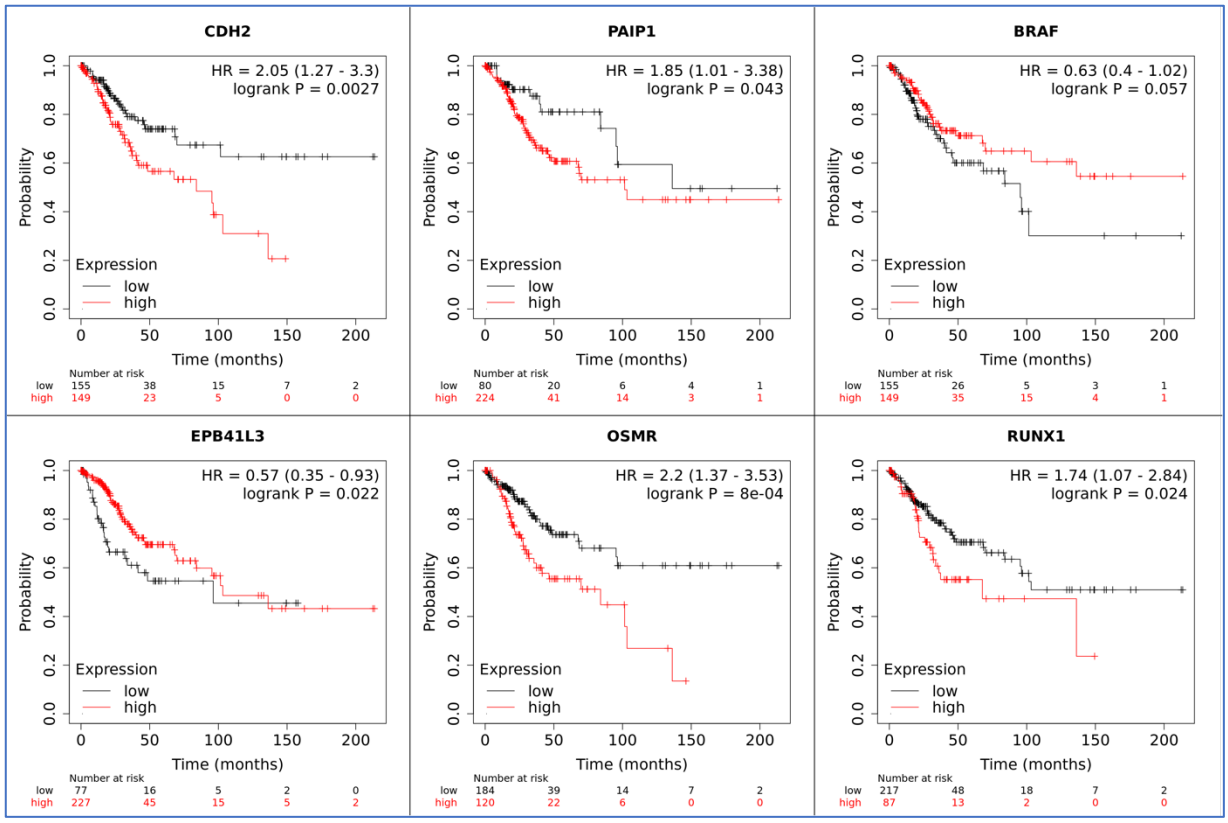


Figure 5: List of genes having poor prognosis in cervical cancer patients as per survival analysis

3.4. Conclusion

The current study focuses on the characterization of multifractal behavior and cross-correlation analysis to prioritize the potential candidate genes involved in cervical cancer by 2D MF-X-DFA in combination with chaos game representation. The study mainly focuses on the frequency CGR matrix generation for each protein sequence and analyzing fractal and cross-correlation behavior. This study prioritizes a total of six genes CDH2, PAIP1, BRAF, EPB41L3, OSMR and RUNX1, which show poor prognostic performance in cervical cancer patients. Further experimental analysis is needed to evaluate the efficacy of the prioritized genes.

Chapter 4

Analysis of protein-protein interaction networks in cervical cancer

4.1. Introduction

Women worldwide suffer high morbidity and mortality rates because of cervical cancer (CC), which holds fourth rank among different types of cancers (Bray et al., 2018 & Sung et al., 2021). It is divided into two subtypes, i.e., squamous cell carcinoma (80%-90%) and adenocarcinoma (10%-20%). India contributes at least one-fourth of the disease burden globally (Reichheld et al., 2020). Human papillomavirus (HPV) is the primary risk factor for CC. However, weak immune system, smoking, birth control pills, and multiple sexual partners also play an imminent role as risk factors (Cohen et al., 2019). It is evident that tumor progression involves various genetic and epigenetic events along with risk factors. Hence, it is important to elucidate the molecular mechanisms involved in tumor progression to understand the disease better. Currently, available treatment options are surgery, radiotherapy, and chemotherapies that don't give protection against the disease, as 75% of CC patients develop further progression or recurrent/recurrence of tumors. Every treatment strategy depends solely on the tumor heterogeneity of the patient (Cook et al., 2011). Understanding the gene expression pattern among the patients is essential to predict diagnostic and prognostic gene signatures that could be used to diminish the outcome of the disease in combination with protein-protein association networks (Oany et al., 2021).

The high-throughput gene expression profiling methods are widely used in cancer genomic studies to understand the molecular classification of the disease, patient stratification, prognosis, and new drug targets (Kulasingam & Diamandis, 2008; Nannini et al., 2009; Bustin & Dorudi, 2004 & Liang et al., 2016). Gene expression profiling methods are known to reveal the differential expression of the genes and their respective dysregulated pathways responsible for the disease progression. Integrating biological knowledge with protein-protein interaction (PPI) networks provides blueprints to understand the complex structural organization of disease-related networks (Chen et al., 2019).

The structural organization of the PPI network consists of nodes and edges representing proteins and their interactions, respectively (Rual et al., 2005). Topological structural analysis of the PPI network reveals the biological significance of each protein in the network (Raman, 2010 & Stelzl et al., 2005). Hence, it is essential to identify crucial proteins responsible for maintaining the global structural stability of the PPI network. Hubs and bottlenecks are

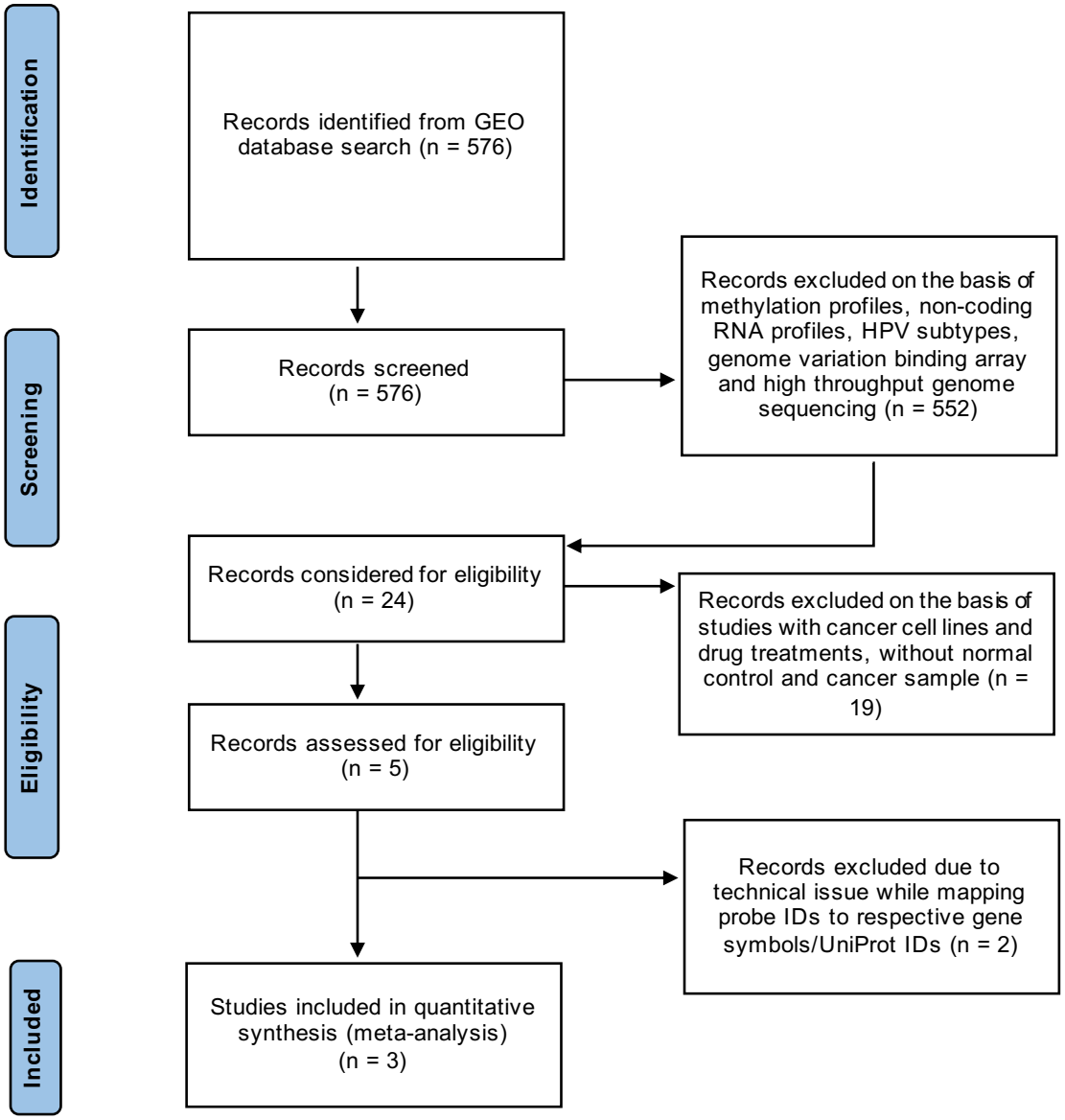
extensively studied in network analysis to understand their importance in the network. Hubs are proteins with a large degree, and bottlenecks are proteins with high betweenness centrality, known to be central molecules in the PPI network (Barabási et al., 2011; Yu et al., 2007 & Ghasemi et al., 2014). This study assessed hubs and bottleneck properties that might help underpin the molecular mechanisms underlying carcinogenesis and develop better intervention strategies. Along with topological centrality calculations, relative vulnerability analysis of a network aims to identify critical proteins that can affect the overall structural stability of the network (Podder et al., 2018).

Our present study combines differential expression analysis, protein-protein interaction network analysis with functional enrichment, and survival analysis to identify the potential key molecular players involved in cervical cancer progression.

4.2. MATERIALS and METHODS

4.2.1 Retrieval and pre-processing of datasets

Cervical cancer gene expression profiles were searched with the keywords "cervical cancer" and "microarray" at the Gene Expression Omnibus, a genomic data repository of the National Center for Biotechnology Information (NCBI, GEO). The inclusion criteria for selecting the datasets were that the study should be focused on cervical cancer as the main subject from the organism *Homo sapiens*. The main aim of the study was to filter the protein-coding genes (mRNA) that were significantly differentially expressed in tumor tissues with respect to the normal tissues. Our query resulted in the identification of 576 studies, and out of these, studies that didn't satisfy inclusion criteria were excluded (**Supplementary Figure 1**). Five datasets eligible to be considered for the analysis were obtained (Zhai et al., 2007; Den Boon et al., 2015; Scotto et al., 2008; Wong et al., 2003 & Travasso et al., 2008), but two datasets were excluded due to a technical issue while mapping probe IDs to gene symbols. Finally, three potential datasets were selected for further analysis. The selected datasets considered in this study are summarized in Table 1.



Supplementary Figure 1: PRISMA flow chart of the microarray meta-analysis for the selection of cervical cancer datasets.

Table 1: The details of the microarray datasets from the NCBI GEO database

S.No	Dataset	Accession No.	Platform	Sample size
1.	Human pre-invasive and invasive cervical squamous cell carcinomas and normal cervical epithelia (Zhai et al., 2007)	GSE7803	GPL96 [HG-U133A]	10 normal and 28 cancer samples
2.	Gene expression analysis of cervical cancer progression (Den Boon et al., 2015)	GSE63514	GPL570 [HG-U133_Plus_2]	24 normal and 28 cancer samples. It also contained 14 CIN1, 22 CIN2 & 40 CIN3 lesions
3.	Identification of gene expression profiles in cervical cancer (Scotto et al., 2008)	GSE9750	GPL96 [HG-U133A]	24 normal and 33 cancer specimens. It also contained samples of 9 cell lines

Our study mainly focuses on screening differentially expressed genes (DEGs) between tumor and adjacent normal samples, and the expression profiles from cell line studies were excluded. Collected datasets were processed using *affy* (Gautier et al., 2004) and *limma* (Ritchie et al., 2015) libraries of the Bioconductor package in the R platform. Microarray datasets are preprocessed as follows: the probe sets in the dataset were normalized through the RMA (Robust Multi-array Average) function of the *affy* package to obtain expression values. The probe sets with expression values were annotated to respective official Gene Symbols. Replicated entries of a gene were removed to reduce the noise, and missing gene expression values across rows and columns were imputed. The processed unique gene expression matrix with its gene symbols was used for the subsequent analysis.

4.2.2 Differential expression analysis

Differential gene expression analysis was carried out on the processed dataset through Linear Models for Microarray Analysis (LIMMA). Genes differentially expressed in cancer samples with respect to normal samples from each dataset were identified using the *limma* package of R (Ritchie et al., 2015). Statistically significant differentially expressed genes (DEG) were filtered based on the criteria of a log fold change > 2 and adjusted P-value < 0.01. The distribution of DEGs in each dataset was visualized through volcano plots. The same procedure is followed for all the datasets except annotation, as different platforms generated them. Furthermore, batch effects in datasets were corrected by computing effect sizes with random-effects models from the metafor R package (Viechtbauer 2010).

4.2.3 Construction of protein-protein interaction network

Identified DEGs were investigated for their interactions through the protein-protein interaction (PPI) network. The PPI network was constructed by retrieving all the available interactions of human from various databases such as APID, DIP, HitPredict, PIP, i2D, BioGrid, MINT, STRING, and IntAct (Alonso-López et al., 2019, 2016; Xenarios et al., 2000; Patil et al., 2011; McDowall et al., 2009; Kotlyar et al., 2016; Oughtred et al., 2021; Licata et al., 2012; Szklarczyk et al., 2021 & Kerrien et al., 2012). Proteins participating in PPIs were mapped to official gene symbols first, then all the interactions were merged. Finally, only those interactions in which both the interacting proteins are part of identified DEGs were extracted. To get the final simplified network, self-loops and duplicate edges were removed from the primary network. The PPI network construction and analysis were carried out using the *igraph* R package (Csardi & Nepusz et al., 2006) to find the essential proteins in the network.

4.2.3.1 PPI network topology analysis

Topological centrality analysis was performed on the disease-related protein-protein interaction network to identify the critical nodes in the network based on hubs and bottleneck properties. As per the degree centrality, each protein in the network was assigned a degree value, then a cutoff was calculated based on the 80-20 rule (Newman 2005). All the proteins above the cutoff were considered hubs. Betweenness centrality measures the number of shortest

paths passing through a particular node. Hence, nodes with the highest betweenness control the network's information flow, representing the network's critical points. These nodes were referred to as the "bottlenecks" of the network. The cutoff was calculated by the same 80-20 rule, and all the proteins having betweenness above the cutoff were considered bottlenecks.

4.2.3.2 PPI network vulnerability analysis

Network vulnerability analysis was performed to identify the most vulnerable proteins as therapeutic targets for cervical cancer. The overall structural stability of the PPI network of cervical cancer was assessed by deleting proteins randomly from the core network and analyzing three topological parameters: Clustering coefficient or transitivity, Average path length (APL), and heterogeneity after each node removal (Podder et al., 2018). The average path length is a measure of the network's overall connectivity. The APL in a network is obtained by calculating the mean of the shortest paths between all pairs of nodes (both ways for directed graphs). The clustering coefficient measures the probability that the adjacent nodes of a particular node are connected. It is also referred to as transitivity. Network heterogeneity is defined as a network with heterogeneously distributed nodes with a higher number of connections, i.e., hubs, as well as a low number of connections. $Heterogeneity = \sqrt{\frac{var(k)}{mean(k)}}$, where k = node degree of the network.

The analysis was carried out by removing one node (protein) from the network and calculating each parameter for the rest of the network. The simulated networks number for each property should be exactly equal to the number of nodes (proteins). Average path length, clustering coefficient, and heterogeneity were estimated in the presence and absence of each protein to assess the influence on the overall network. With this analysis, we can figure out the most valuable proteins that are critical for maintaining the structural stability of the network. The most important ones will be the outliers in the graphs generated for the parameters. The outliers indicate that removing that particular protein affects the whole network. Hence, it is crucial and can be considered a drug target by analyzing and integrating the outcome of the vulnerability and hub-bottleneck analyses.

4.2.4 Functional enrichment analysis

The combined list of proteins obtained by vulnerability and topology analysis of disease-related networks are assessed for their associated role in various processes and pathways. Gene Ontology (GO) terms and Kyoto Encyclopedia of Genes and Genomes (KEGG) pathways for the respective proteins were assessed through the *clusterProfiler* R package (Yu et al., 2012). GO is mainly used in functionally annotating genes based on GO terms: Biological process (BP), Molecular function (MF), and Cellular component (CC). KEGG, a pathway database, allows us to explore the associated pathways for the given set of proteins.

4.2.5 Validation of gene expression at the protein level

Immunohistochemistry data available at Human Protein Atlas (HPA), a human proteome map (Uhlén et al., 2015), is accessed to validate the key gene's protein expression patterns based on the staining intensity levels in both normal cervix tissue and cervical cancer tissues.

4.2.6 Survival analysis

To further characterize the candidate genes, survival analysis was performed through the Kaplan-Meier survival curve and log-rank test through the Kaplan-Meier plotter online tool (Lánczky & Győrffy 2021). Overall survival (OS) of patients depends on the time between the surgery date and death or the last follow-up date. The GEO datasets don't have information related to clinical profiles. The clinical data was accessed from the TCGA-CESC project on cervical cancer to investigate the prognostic performance of the candidate genes in patients and p-values <0.05 are considered statistically significant prognostic factors for cervical cancer.

4.3 Results

4.3.1 Identification of differentially expressed genes between different tissue samples

In the present study, differential expression analysis identified 544 unique genes as differentially expressed between cancer and normal tissue samples from three datasets (GSE7803, GSE63514 and GSE9750). Out of 544 differentially expressed genes, 248 were upregulated and 296 were downregulated. The datasets GSE527 and GSE4482 were not considered for further studies due to a technical issue while mapping probe IDs to respective

gene symbols/UniProt IDs. The detailed distribution of DEGs in each dataset is given in Table 2 and represented as volcano plots (Figure 1).

Table 2: No. of DEGs present in each dataset through differential expression analysis

GEO accession	No. of genes after processing	No. of DEGs	Upregulated genes	Downregulated genes
GSE7803	12403	79	21	58
GSE63514	12403	412	211	201
GSE9750	12403	265	61	204

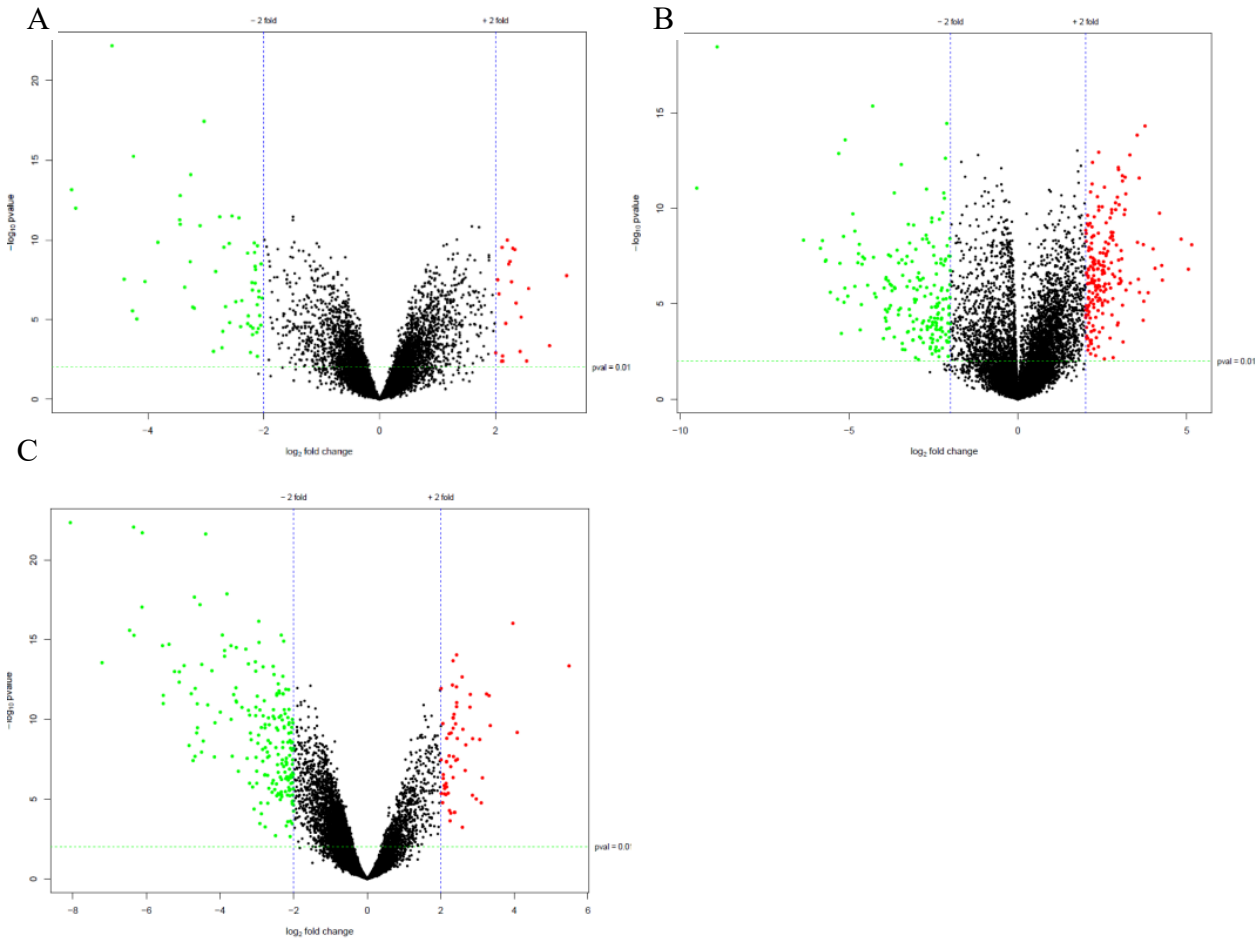


Figure 1: Number of DEGs in each dataset. (A) GSE7803, (B) GSE63514, (C) GSE9750. Green dots are downregulated genes and Red dots are upregulated genes. Cutoff: logFC= 2 & pval= 0.01.

4.3.2 Essential proteins in the dysregulated network of cervical cancer were identified through network analysis

After merging all the protein-protein interactions from different PPI databases, there were 14,30,022 interactions among 51,209 proteins. After removing redundancy, we got 8,52,432 interactions among 51,209 proteins. Then, only those interactions in which both the proteins were part of the DEGs list were extracted. Thus, the number of DEGs interactions were 4,942 with 498 proteins. The network was visualized using Cytoscape (Shannon et al., 2003) for further analysis (Table 3) (Figure 2).

Table 3: Network analysis summary

Properties of Networks	No.
Number of nodes	498
Number of edges	4942
Avg. number of neighbors	20.114
Network diameter	8
Network radius	5
Characteristic path length	3.160
Clustering coefficient	0.378
Network density	0.041
Network heterogeneity	1.216
Network centralization	0.176
Connected components	4

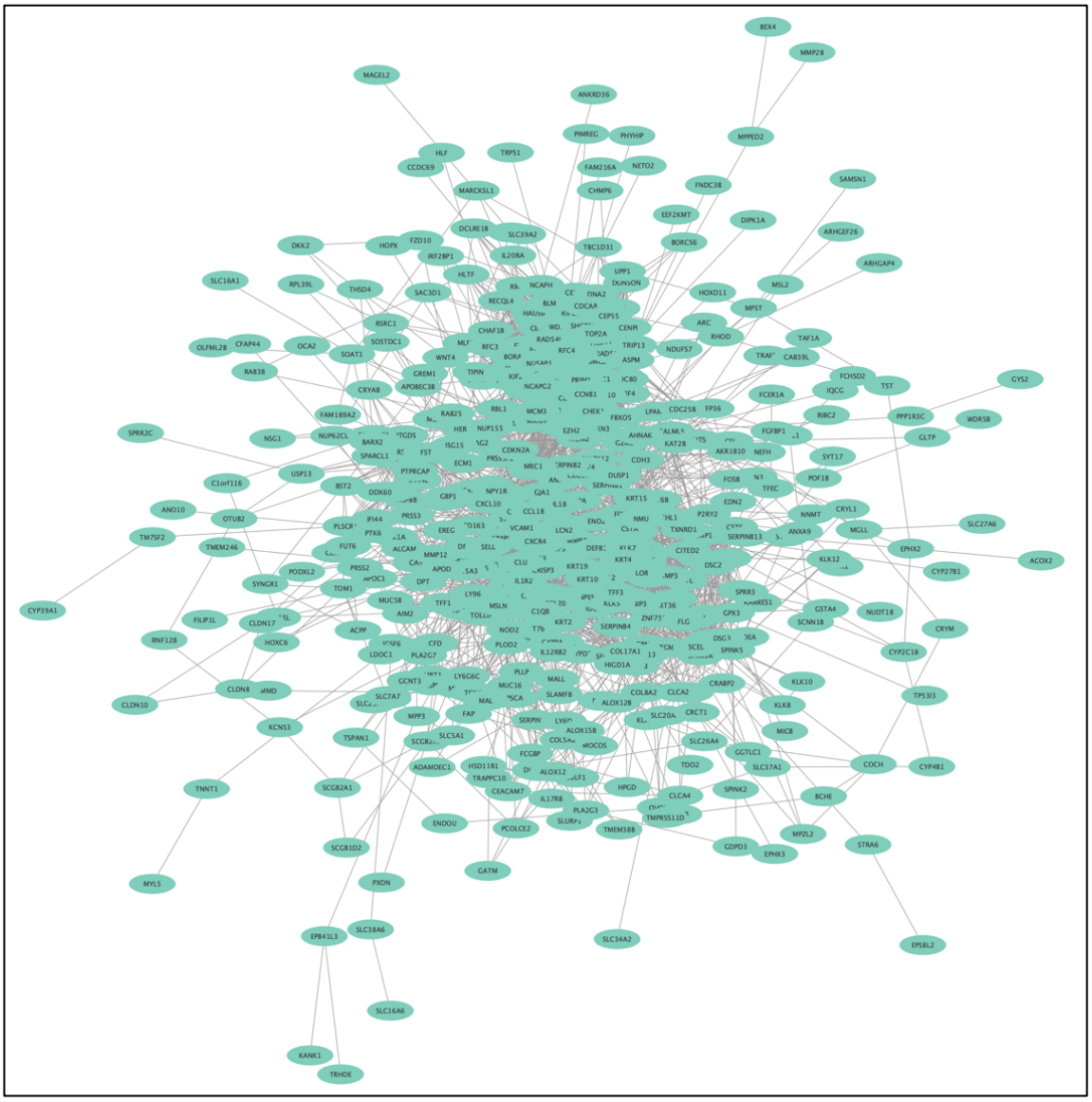


Figure 2: Protein-protein interaction network of DEGs representing 4942 interactions and 498 nodes/proteins.

All proteins with a degree value of > 63.62 were considered hubs following the 80-20 rule. It resulted in 55 proteins as hubs. Further, all proteins with a betweenness value > 0.02509 were considered bottlenecks using the same 80-20 rule, resulting in 14 bottlenecks. Seven proteins were found to have properties of both hub and bottleneck (Table 4) (Figure 3).

Table 4: List of key proteins resulted from network analysis

Hub proteins	MCM2, CDK1, TOP2A, KIF11, CCNB1, MKI67, CDC6, MCM5, CHEK1, MCM10, NDC80, TTK, BUB1B, CDC45, MCM6, EXO1, FN1, CXCL8, RFC4, MCM3, KNTC1, TRIP13, ASPM, MELK, FANCI, RAD51AP1, KIF2C, DTL, OIP5, SMC2, CENPE, KIF23, NCAPH, DLGAP5, CDCA8, KIF15, WDHD1, KIF20A, KIF4A, CEP55, NUSAP1, PRC1, RAD54L, GINS2, POLE2, CDKN3, HMMR, FOXM1, MMP9, PRIM1, SPAG5, HELLS, NCAPG2, EZH2, and FBXO5
Bottleneck proteins	TRIM16, FN1, MCM2, CXCL8, MMP9, MKI67, CCND1, TRIP13, FOS, ISG15, VCAM1, MCM5, IGF1, AGR2
Common Hub and Bottleneck proteins	MCM2, MKI67, MCM5, FN1, CXCL8, TRIP13 and MMP9
List of Vulnerable proteins	MCM2, MKI67, KIF11, CCNB1, CDC6, TTK, CDC45, BUB1B
Potential key genes	MCM5, FN1, TRIP13, KIF11, TTK, CDC45, and BUB1B

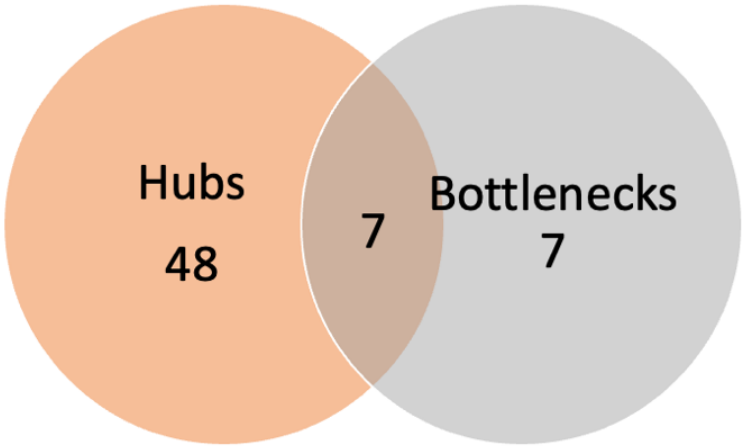


Figure 3: Venn diagram illustrating the number of common hubs and bottlenecks

4.3.2.1 Network vulnerability analysis identifies critical proteins with the property of altering the structural stability of the network

Network vulnerability analysis was performed to assess the three topological properties such as APL, transitivity, and heterogeneity, which were calculated for the simulated network after deleting one node at once using a scatter plot, and outliers were identified to pinpoint the important proteins as potential drug targets for cervical cancer (Table 5) (Figure 4).

Table 5: Network vulnerability analysis

Gene name	Average path length	Gene name	Transitivity	Gene name	Heterogeneity
MCM2	3.197112	CDC45	0.658095124	MCM2	5.379136753
FN1	3.188606	TTK	0.658436214	CDK1	5.388003605
CXCL8	3.185894	EXO1	0.658443901	TRIP13	5.388986277
MKI67	3.17975	KIF20A	0.658578871	TOP2A	5.391699113
MMP9	3.177497	KIF23	0.658595808	CCNB1	5.393047379
ISG15	3.176812	ASPM	0.65875306	CDC6	5.395715251
AGR2	3.174701	DTL	0.658784218	CHEK1	5.396180819
IGF1	3.174358	CDCA8	0.658894051	KIF11	5.396506068
FOS	3.173623	BUB1B	0.65894888	MCM10	5.397525741
CCND1	3.17349	KIF15	0.658949144	MCM6	5.397692815
KIF11	3.170953	MELK	0.658972469	MKI67	5.39817741
EZH2	3.1696	RAD51AP1	0.659088476	TTK	5.398372616
MCM5	3.168432	CDC6	0.659095901	CDC45	5.398784435
VCAM1	3.168157	KIF4A	0.65921154	BUB1B	5.399276182
NDC80	3.168006	CCNB1	0.659270179	RFC4	5.400048301

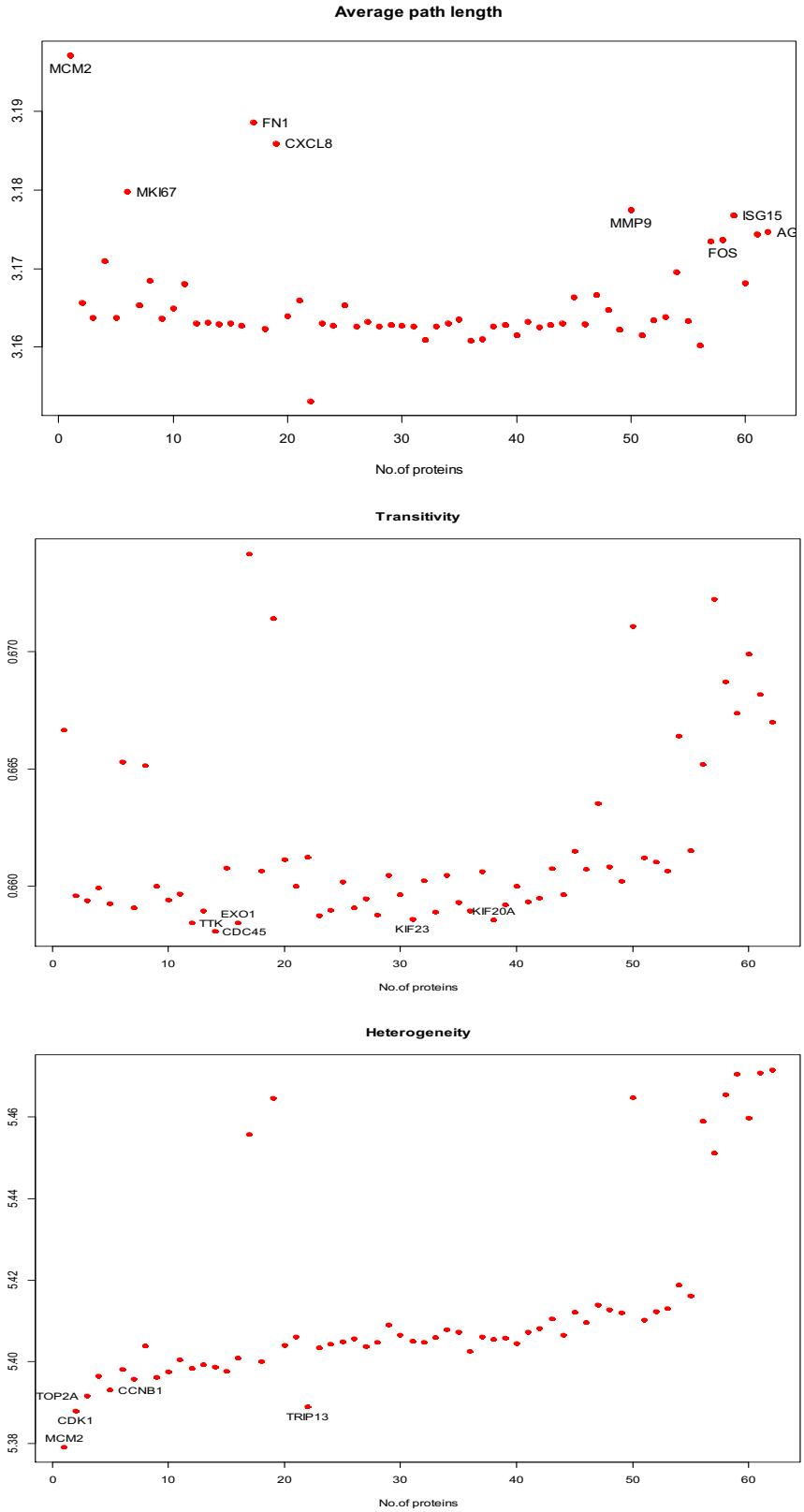


Figure 4: Topological vulnerability assessment of cervical cancer network

Outliers are the most critical proteins of the network, which can initiate disease progression by changing the stability of the PPI network. Common outliers were identified from the three graphs, then common outliers from any two graphs taken at a time. The list of vulnerable proteins was found to be MCM2, MKI67, KIF11, CCNB1, CDC6, TTK, CDC45, and BUB1B. The list of the common hub and bottleneck proteins are found to be MCM2, MKI67, MCM5, FN1, CXCL8, TRIP13, and MMP9. Both vulnerable and common hub and bottleneck proteins were considered to prioritize potential candidates for the disease. Further, The reports of their involvement in cervical carcinogenesis were further validated by us in the Gene Cards database under MalaCards (Rappaport et al., 2017). It was found that MCM5, FN1, TRIP13, KIF11, TTK, CDC45, and BUB1B were not reported for their involvement in the disease, and the rest were reported for their involvement in cervical cancer.

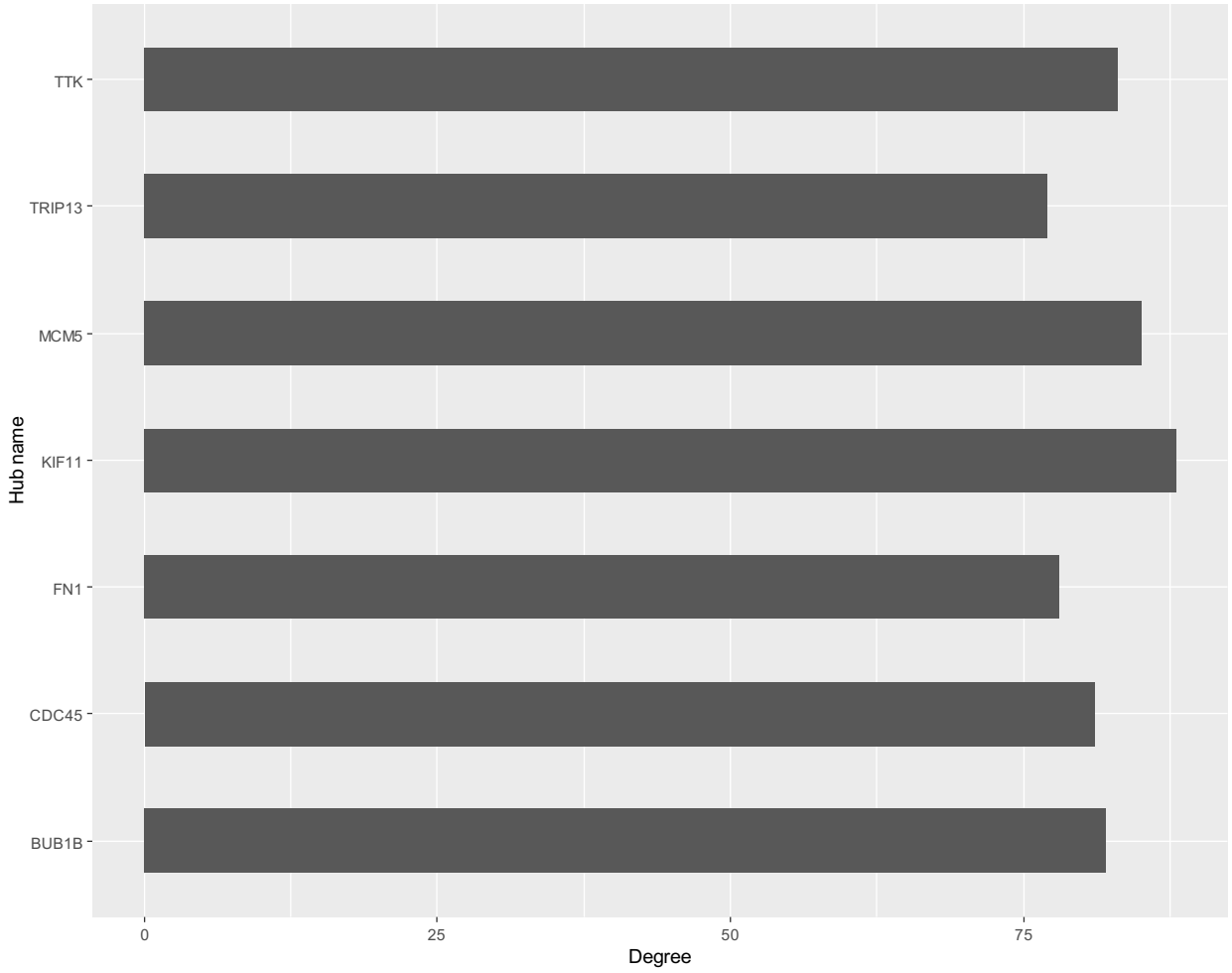


Figure 5: Candidate genes with their corresponding degree centrality values.

Further analysis was proceeded with these seven proteins as novel drug targets. Their corresponding degree centrality was represented as a bar plot (Figure 5). The key gene PPI network was visualized using Cytoscape (Figure 6).

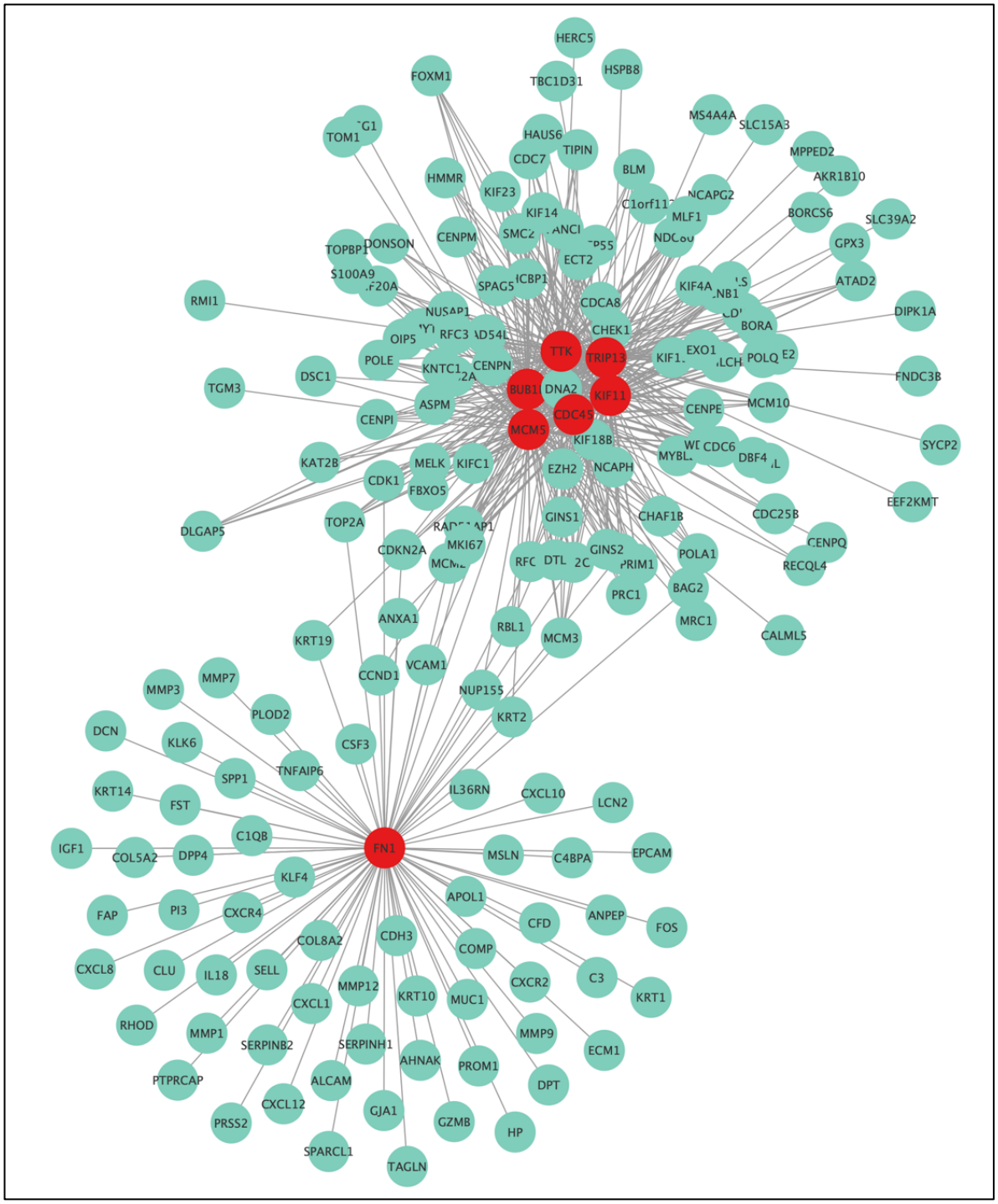
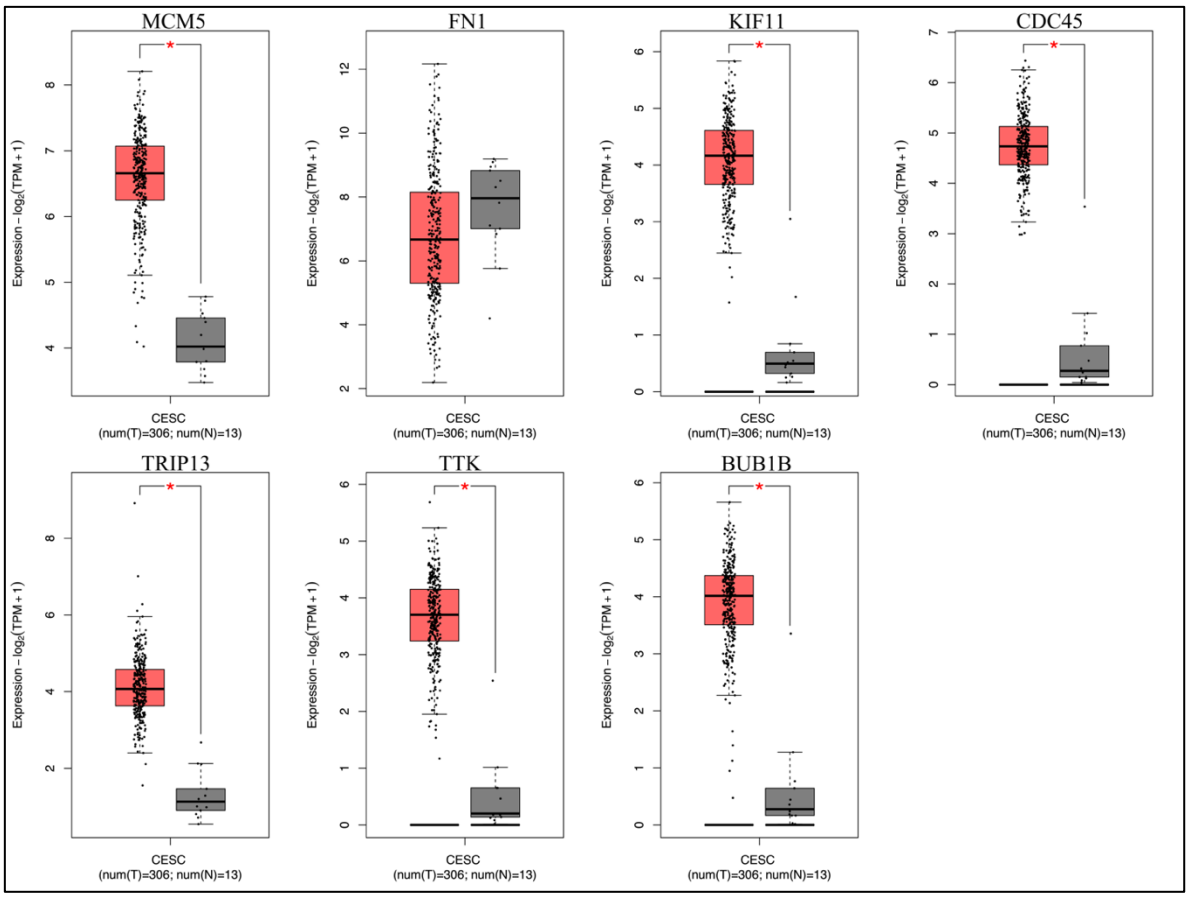


Figure 6: PPI network of key genes with two clusters.

The key genes identified from the microarray data analysis were verified with NGS data of cervical squamous cell carcinoma available at the TCGA data portal. Key genes identified in the study followed the same patterns of gene expression in NGS data also visualized as box plots (Supplementary Figure 2).



Supplementary Figure 2*: NGS data analysis: Relative expression profiles of key genes between tumor vs normal tissues of CC patients visualized as box plots (red colour = tumor & black colour = normal samples). It represents that their expression follows the same pattern in both microarray and NGS data.*

4.3.3 Essential candidate genes were associated with cell-cycle-related GO terms and KEGG pathways

Gene ontology and KEGG pathway enrichment analysis of the seven key genes were investigated using the *clusterProfiler* R package. The proteins were mostly present in biological processes related to cell proliferation, cell division, cell cycle checkpoint, and

mitotic nuclear division. Also, the proteins were mainly involved in the cell cycle related KEGG pathways (Figure 7). p-value <0.05 were considered as statistically significant GO terms, and KEGG pathways and p-values were adjusted by Benjamini–the Hochberg method.

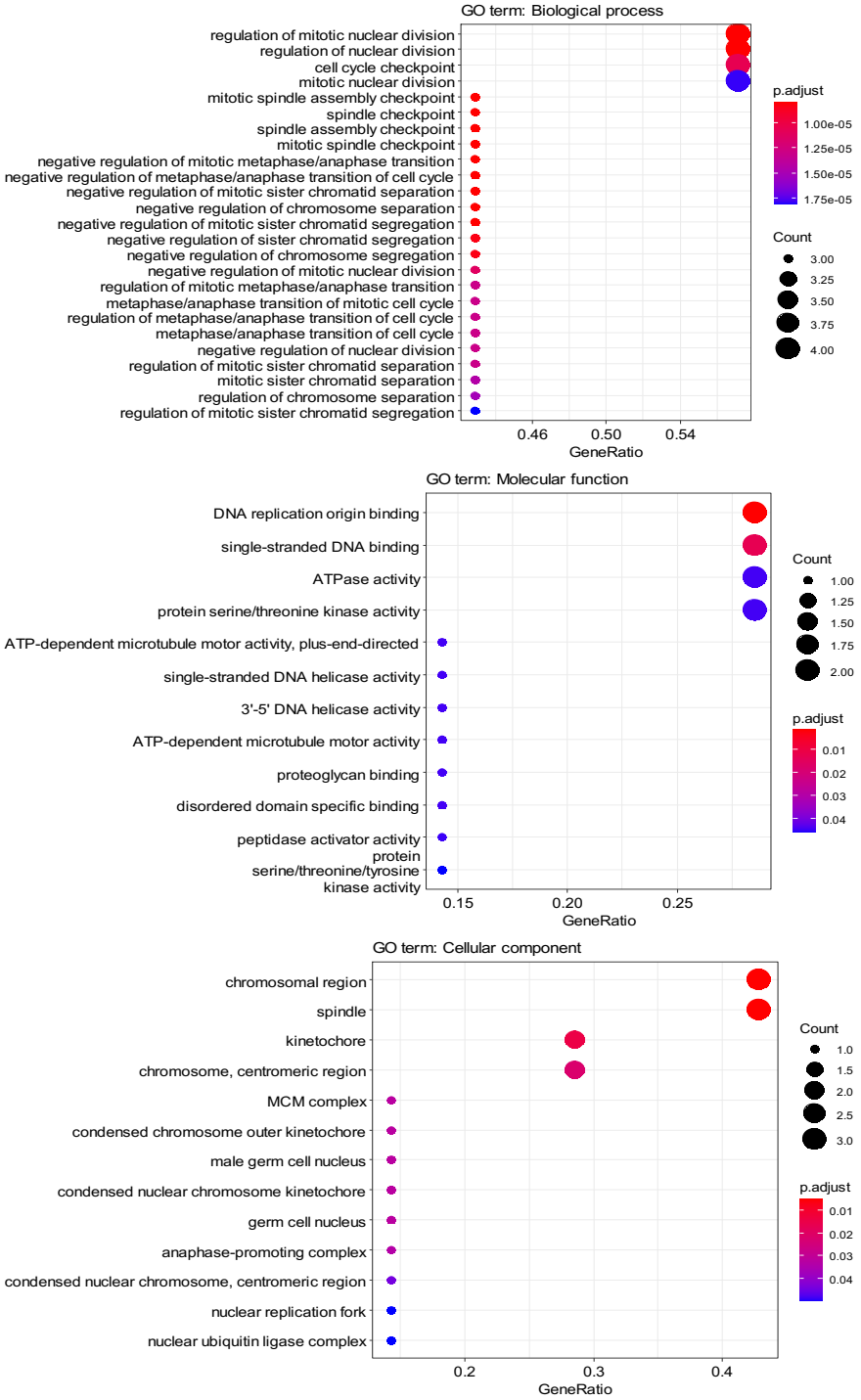


Figure 7: Gene Ontology (GO) analysis of prioritized potential candidate genes (pvalue<0.05)

4.3.4 Validation of protein expression levels of key candidate genes via immunohistochemical data

The protein expression levels of the candidate proteins were also verified using the immunohistochemistry mapping data in both normal and cancer tissues via the HPA database. The proteins MCM5, FN1 and KIF11 were highly detected with strong intensity levels in cervical cancer tissues. In normal cervix tissues, MCM5 has shown medium expression in squamous epithelial cells and was not detected in glandular cells. FN1 is not detected in normal tissue. KIF11 has high expression in glandular cells and medium expression in squamous epithelial cells (Supplementary Figure 3a). Likewise, CDC45, TRIP13, and TTK proteins have medium expression levels with moderate intensity in cancer tissues. CDC45 and TRIP13 were not detected in glandular cells and had low expression in squamous epithelial cells of normal cervix tissue. TTK has low expression in both glandular and squamous epithelial cells (Supplementary Figure 3b). The BUB1B protein expression data was unavailable in the HPA database for both normal cervix and tumor tissues. From these results, it can be observed that higher protein expression patterns were significantly related to the prognosis of cervical cancer patients and can be explored as potential candidate molecules for the patient's survival.

4.3.5 Prognostic performance of the candidate genes shows that their expression negatively correlated with overall survival

Kaplan–Meier survival curve and log-rank test assessed the candidate genes prognostic significance in cervical cancer patients. Cervical cancer data available at TCGA contains 304 patient samples with clinical information. The cancer patient's samples were grouped into high and low groups on the basis of median expression values. Overall survival of patients depends on the expression levels of the respective prognostic candidate genes.

It was observed that candidate gene expression levels were negatively correlated with overall survival with a statistical significance of p-vlaue< 0.05. Higher expression of the FN1 gene and lower expression of the MCM5, KIF11, and CDC45 genes have poor prognosis in CC patients (Figure 8).

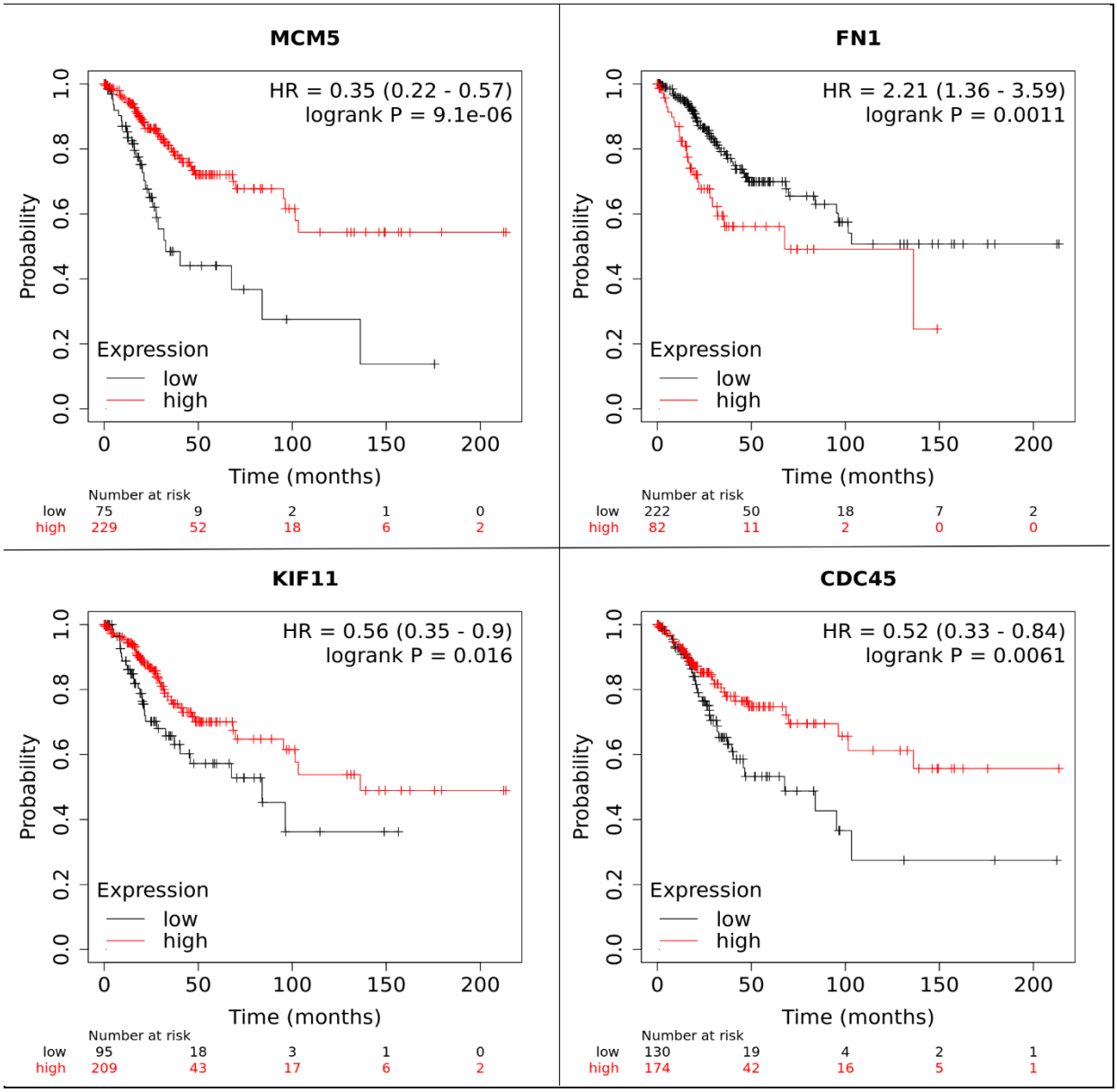


Figure 8: Prognostic significance of candidate genes in cervical cancer patients. Kaplan-Meier plot representing overall survival analysis of survival difference between high and low-risk cervical cancer patients from the TCGA datasets. The X-axis represents overall survival time in months and the y-axis represents the probability of survival. Pvalues<0.05 were considered statistically significant prognostic factors for cervical cancer.

4.4Discussion

Cancer is a heterogeneous and genetic disease involving a series of alterations in the genetic makeup, causing malignant transformation in disease progression (Yu & Henneberg 2018).

The present study focuses on cervical squamous cell carcinoma, a subtype of cervical cancer, which holds the fourth rank with other cancers among women all over the world. To understand the genetic changes occurring during the progression of cervical carcinoma and to identify novel drug targets, we used available bioinformatics methods and in-house code to extract meaningful information from the microarray gene expression profiles and disease-related protein-protein interaction network.

Our findings related to the candidate genes involved in cervical squamous cell carcinoma progression allow experimental researchers to investigate their role as a critical gene in the targeted therapy. Various studies reported extensive utilization of complex network properties to understand the role of key molecules in multiple diseases. Analysis of complex network parameters in the biological networks were found to have the edge over other available methods for identifying candidate genes. An integrative study was performed to propose key molecules using differential expression, protein-protein interaction network, and functional enrichment analysis.

In the present study, differential expression analysis was performed on NCBI GEO datasets (GSE7803, GSE63514, and GSE9750) of cervical cancer to identify significantly expressed genes in the tumor samples with their matched normal samples. Statistical filtering identifies a total of 544 genes as differentially expressed genes (DEG). Among 544 genes, 248 were found to be upregulated, and 296 were downregulated genes. Further, protein-protein interactions (PPI) of significantly differentially expressed genes were extracted by mapping DEGs on the reconstructed human protein-protein interaction network. We performed network analysis and calculated complex network parameters using in-house code to filter the most valuable proteins in the PPI network that can be used as a biomarker for the disease.

The key gene PPI network consists of 174 nodes and 558 edges with two clusters that represent molecular complexes in the PPI network. FN1 protein with 79 nodes and 78 edges forms one cluster and the rest six proteins (MCM5, TRIP13, KIF11, TTK, CDC45 and BUB1B) form another cluster with 119 nodes and 481 edges in the network. These two clusters were densely connected in their respective networks with each other by interacting with other proteins which are part of both the clusters, such as MCM6, MKI67, KIF2C, ANXA1, RBL1, RFC4, BAG2, MCM3, MCM2, CDKN2A, CDK1, TOP2A, MCM2, KRT19, VCAM1, CCND1, NUP155 and

KRT2. This set of proteins acts as linkers between the two key protein clusters to maintain structural integrity in the protein-protein interaction network related to cervical squamous cell carcinoma. KEGG pathway analysis reports for these proteins were associated with cell cycle, DNA replication, cellular senescence, viral carcinogenesis, platinum drug resistance and the p53 signaling pathway. Also, highly enriched in the gene ontology terms of biological processes such as the G1/S transition of the mitotic cell cycle, DNA replication, DNA conformational change, DNA duplex unwinding, regulation of myeloid cell apoptotic process, etc.

MCM5 (Minichromosome Maintenance Complex Component 5) is part of the MCM family and plays a vital role in the initiation of DNA replication. MCM5 is associated with several cancers, such as breast cancer (Eissa et al., 2015), ovarian cancer (Levidou et al., 2012), oral squamous cell carcinoma (Yu et al., 2014), etc.; also, few studies highlighted it's role in cervical cancer tumor progression (Qing et al., 2017 & Li et al., 2018). FN1 (Fibronectin 1), a glycoprotein belonging to the FN family, plays various cellular activities such as cell migration, cell adhesion, and cytoskeletal organization in multiple diseases (Pankov & Yamada 2002; Mao et al., 2005 & Gao et al., 2016). In various tumors, such as osteosarcoma, nasopharyngeal carcinoma, esophageal cancer, and ovarian cancer, FN1 is a critical tumor-related gene (Jiang et al., 2017; Song M et al., 2017; Song G et al., 2017 & Lou et al., 2013). TRIP13 (Thyroid Hormone Receptor Interactor 13) is associated with the AAA (ATPase family associated with various cellular activities) protein superfamily and plays crucial roles in regulating various cellular processes such as chromosome synapsis, DNA break repair and recombination, and checkpoint signaling (Miniowitz-Shemtov et al., 2015 & Vader 2015). TRIP13 is one of the critical genes, acting as a tumor susceptibility locus, related to Chromosome instability (CIN) in human tumors and is associated with poor survival in various tumors (Zhou et al., 2013; Wang et al., 2014; Yost et al., 2017; Carter et al., 2006 & Lu et al., 2019).

KIF11 (Kinesin Family Member 11) is a motor protein essential for spindle dynamics, including centromere separation, chromosome positioning, and bipolar spindle establishment during mitosis (Rapley et al., 2008 & Ferenz et al., 2010). Previous reports suggest that KIF11 is associated with lung cancer, glioblastoma, malignant mesothelioma, and gastric cancer (Schneider et al., 2017; Venere et al., 2015; Kato et al., 2016; Imai et al., 2016 & Daigo et al.,

2018). TTK (Threonine and Tyrosine Kinase), also known as human monopolar spindle 1 (HMPS1), a mitotic protein kinase, plays a crucial role in regulating cell division via mitotic checkpoints and chromosome attachment. Overexpression of TTK affects chromosomal instability, further resulting in tumor progression (Benzi et al., 2020; Silva et al., 2018 & Lim et al., 2017). It is associated with various cancers, such as breast cancer, glioblastoma, thyroid cancer, and is a potential candidate for gastric cancer (Huang et al., 2020 & Kaistha et al., 2014). CDC45 (Cell Division Cycle 45) belongs to the multiprotein complex along with Cdc6/Cdc18 and DNA polymerase, which are crucial for the initiation of eukaryotic DNA replication (Masai et al., 2005). Earlier research found that CDC45 is an antigen that promotes cell growth and is linked to the development of cancerous tumors (Pollok et al., 2007 & He et al., 2021).

Furthermore, it is involved in cervical cancer prognosis, which indicates the reliability of our findings (Qiu et al., 2020). BUB1B (BUB1 Mitotic Checkpoint Serine/Threonine Kinase B) is associated with the spindle assembly checkpoint family member of proteins. BUB1B is associated with various biological processes, including chromosome segregation, differentiation of post-mitotic neurons, DNA repair, and ciliogenesis. Previous reports indicated that this protein has a crucial role in tumor progression and prognosis in multiple cancers (Sekino et al., 2021). In addition to the findings from the study, all the potential prognostic candidate genes can be subjected to experimental studies to understand their role in the progression of cervical carcinoma.

4.5 Conclusion

To summarize, systems biology methods were applied to microarray data of cervical cancer and key genes that are involved in the progression and survival of the disease were identified. These genes are MCM5, FN1, TRIP13, KIF11, TTK, CDC45, and BUB1B. They are associated with cell cycle regulation, extracellular matrix remodeling, and chromosome segregation. These genes have prognostic significance for cervical cancer patients, as they can predict the outcome and response to treatment. These genes could be potential targets for developing new treatments for the disease, as they can modulate the biological processes that are altered in cervical cancer cells.

Chapter 5

Analysis of integrative networks to uncover regulatory elements associated with cervical cancer progression

5.1 Introduction

Interactome networks are graphs of the physical interactions among cellular components, such as proteins, DNA, RNA, metabolites, and drugs. These interactions are measured using several techniques and aggregated into a single network that demonstrates the dynamics and complexity of the biological system. Interactome networks facilitate our understanding of the execution, regulation, and coordination of biological functions by many molecular entities, as well as the ways in which genetic or environmental changes impact these processes (Vidal et al., 2011 & Zanzoni et al., 2009).

Interactome data can be of different types, including protein-protein and protein-DNA interactions, metabolic reactions, and signaling pathways. Interactome networks have some common patterns and principles, such as scale-free distribution, small-world phenomenon, network motifs, and network evolution. Interactome networks can be used to study disease phenotypes, such as disease modules, disease genes, disease pathways, network-based stratification, network perturbations, and network-based drug discovery (Caldera 2017 & Yeger-Lotem 2015).

Understanding the gene expression pattern among the patients is essential to predict diagnostic and prognostic gene signatures that can be used to diminish the outcome of the disease in combination with interactome networks (Oany AR et al., 2021). However, interactions between coding and non-coding RNAs affected by microRNAs are also a part of the intricate process of gene regulation, including transcription and post-transcriptional processes. These interactions form protein-protein interaction and gene regulatory networks that affect various cellular processes and influence cancer development and response to therapy. Gene regulatory networks influence development of cancer by changing the oncogenes and tumor suppressors expression, promoting or inhibiting cell proliferation, survival, differentiation, and migration, and modulating the tumor microenvironment and immune response.

A novel integrative networks approach was employed to construct a cancer-specific gene regulatory network and to identify unique genes and sub-networks that are enriched or depleted in certain network motifs and hub proteins. These genes may have therapeutic potential for cancer treatment, but their interactions with tumor cells and stromal cells need to be better

understood. In interaction networks, centrality metrics are widely used to find the most influential nodes. Different centrality measures quantify various aspects of node importance, such as degree, closeness, betweenness, and eigenvector centrality. Hubs and bottlenecks are key nodes in interaction networks. Hubs have a high degree centrality, while bottlenecks have high betweenness centrality. They influence cellular processes and pathologies such as oncogenesis, diabetes mellitus, and neurodegeneration (Barabási and Oltvai, 2004). For example, hubs modulate gene transcription, signal transduction, and metabolic flux, and their alterations can induce aberrant cell proliferation, survival, and differentiation. Bottlenecks facilitate information transfer between network modules or communities, and their perturbation can compromise cellular coordination and regulation (Jeong et al., 2001). Therefore, identifying and targeting hubs and bottlenecks can provide novel insights and strategies for disease diagnosis, prognosis, and therapy.

Next-generation sequencing (NGS) technology has revolutionized our understanding of the human genome, revealing its remarkable complexity and diversity. Among the various types of RNA molecules that are transcribed from the genome, ncRNAs constitute a huge and functionally heterogenous group. ncRNAs include miRNAs and lncRNAs, which are 98% of the total RNA in the cell and play crucial roles in various biological processes (Baltimore, D. 2001). ncRNAs regulate genes and are important for cancer research as biomarkers and targets. One of the ways that lncRNAs can be used as biomarkers is by measuring their expression levels in different tissues or biological fluids, such as blood, urine, or saliva. For instance, lncRNA HOTAIR is upregulated in multiple cancers and can be detected in plasma samples of cancer patients (Hajjari, M., & Salavaty, A. 2015). Another way that lncRNAs can be used as biomarkers is by analyzing their interactions with other molecules, such as miRNAs, mRNA, or DNA. For instance, lncRNA MALAT1 can bind to miR-200 family members and modulate their activity in breast cancer (Jo, Hyein et al., 2022). Circular RNAs (circRNAs), long ncRNAs (lncRNAs) and pseudo-genes (Ψ-genes) can regulate messenger RNAs (mRNAs) in different ways. Some of them form RNA–RNA and lncRNA-RNA complexes that affect transcription in the nucleus, while others increase mRNA stability in the cytoplasm.

Moreover, the competing endogenous RNA (ceRNA) mechanism involves both coding and non-coding RNAs that interact with microRNAs (miRNAs) (Saleembhasha & Mishra 2019;

Ala 2020). The ceRNA hypothesis states that ncRNAs that contain miRNA response elements (MREs), such as lncRNAs, circular RNAs (cirRNAs), and some pseudogenes, can act as ceRNAs to sequester miRNAs and modulate their target mRNAs (Salmena L et al., 2011). By competing for a limited pool of miRNAs, lncRNAs can influence gene expression at the post-transcriptional level. miRNAs bind to the 3'-UTRs of target mRNAs and regulate gene expression by causing mRNA degradation or translational repression. These were implicated in multiple stages of cancer development and progression, such as cell proliferation, invasion, apoptosis, metastasis, and angiogenesis (Bartel 2004). Numerous studies have supported this hypothesis and demonstrated that lncRNAs, mRNAs, and other RNAs can act as natural miRNA sponges and influence the expression of multiple target genes (Chen, W et al., 2019).

This study aimed to understand the complex regulatory interactions among protein-coding and non-coding entities in cervical cancer by constructing gene expression-based cancer-specific regulatory networks and protein-protein interaction networks. The topological studies identified key molecules and pathways associated with the disease's progression.

Materials and Methods

5.1.1 Differential expression analysis

Differential expression analysis on the gene expression dataset of cervical squamous cell carcinoma was performed to filter differentially expressed genes (DEmRNAs), miRNAs(DEmiRNAs) and lncRNAs (DElncRNAs). The gene expression dataset of cervical squamous cell carcinoma was dowloaded from The Cancer Genome Atlas (TCGA-CESC). This dataset contains 304 tumor samples and 3 normal samples. Additionally, 19 samples of normal cervical gene expression data were retrieved from the Genotype-Tissue Expression (GTEx) data portal. Duplicate or irrelevant data from the dataset was removed during pre-processing of the dataset. Differential expression analysis was done using the DESeq2 R package (Love et al., 2014). Significantly differentially expressed mRNAs were screened after removing genes with low read counts among cervical tumor and normal tissue samples. Further, the above process was performed to filter DEmiRNAs as well as DElncRNAs.

5.1.2 Gene co-expression network construction and Identification of co-expressed modules

Significantly differentially expressed genes (DE-mRNA) were used to construct a weighted gene co-expression network using the WGCNA R package to find genes with strong correlations across the samples (Langfelder & Horvath, 2008). Quality control was performed on the expression data and outlier samples were removed before proceeding to the detection of modules. Later, a scale-free adjacency matrix that only retained strong correlations, ignoring weak ones, was obtained by considering a soft threshold power of 6. The adjacency matrix was transformed into a topological overlap matrix (TOM). Co-expressed gene modules were detected by applying hierarchical clustering and dynamic tree cut using the dissimilarity of module eigengenes (1-TOM) as a distance measure. The minimum module size was set to 30 and modules with similar expression profiles were merged using a threshold of 0.25 through hierarchical clustering of module eigengenes.

5.1.3 Protein-protein interaction network analysis

To reconstruct the PPI network, all the available interactions of human from various databases such as APID, DIP, HitPredict, PIP, i2D, BioGrid, MINT, STRING, and IntAct were retrieved. The proteins involved in PPIs were mapped to official gene symbols first, then the interactions were merged. Next, only the interactions where both the interacting proteins were part of identified DEGs were extracted. Self-loops and duplicate edges were removed from the primary network to get the final simplified network.

PPI network of both dysregulated genes and significantly correlated co-expressed module with clinical feature HPV status were extracted from the reconstructed human PPI network. To identify critical genes in the dysregulated PPI networks, i.e., core PPI network and module-specific PPI network, network topological structure analysis was performed. Based on the network's degree and betweenness centrality properties, the most important genes that influence its structure and function were selected.

Additionally, the topological features of the core PPIN and co-expressed module-related PPIN were evaluated using Pearson correlation. Further, to identify highly influential genes in the network, the 80/20 rule or Pareto principle was employed. The pareto principle states that the

top 20% of the average value is responsible for 80% of the population's actions. The PPI network construction and analysis were carried out using the igraph R package and visualized with Cytoscape (Shannon et al., 2003).

5.1.4 Module – clinical feature associations

A correlation analysis was conducted to determine the relationship between the co-expressed modules and the clinical characteristics of cervical cancer (CC) patients, using various clinical features as traits. The clinical features included HPV status, age at diagnosis, clinical stage, race, neoplasm histologic grade, menopause status, pathology T/N/M stages and lymphovascular involvement. A stringent filter was applied to select the modules with a high correlation and a low p-value with the clinical features. The biological significance of these modules was further investigated.

5.1.5 Integrative regulatory network construction and analysis

5.1.5.1 lncRNA-mRNA interaction prediction and regulatory network analysis

The regulatory roles of the differentially expressed lncRNAs in CESC patients were investigated by performing lncRNA–target prediction using the LncTarD database (Zhao et al., 2023). This database contains 8,360 experimentally validated lncRNA–target interactions across 419 disease subtypes and their clinical implications.

The predicted interactions were filtered to retain only those involving both differentially expressed elements (lncRNAs and mRNAs) in CESC patients. A CESC-specific dysregulated lncRNA-mRNA co-expression network was constructed on the basis of these interactions and its topological properties were analyzed using the igraph R package. The degree centrality and betweenness centrality of each node, which are measures of how connected and influential a node is in the network, were calculated to find the critical molecular players in the network. The nodes with the highest values of these metrics were selected as the key molecular players in the network. The network with key molecular players was also visualized using Cytoscape.

The potential binding sites of the significantly differentially expressed mRNAs and lncRNAs were predicted using the human lncRNA-mRNA interaction database (http://rtools.cbrc.jp/cgi-bin/RNARNA/index.pl) (Terai et al., 2016). SUMENERGY plots were generated using R.

5.1.5.2 lncRNA-miRNA-mRNA ceRNA network construction and analysis

The function of lncRNA as ceRNA in gene regulation was elucidated by performing target prediction for the differentially expressed molecular entities. The lncRNA-miRNA interaction pairs were obtained from the miRcode database, which contains interactions that are experimentally validated and computationally predicted (Jeggari et al., 2012). The differentially expressed miRNAs (DEmiRNAs) target genes were predicted by integrating the interaction data from three miRNA target databases: miRDB, miRTarBase, and TargetScan (Chen & Wang 2020; Huang et al., 2021; McGeary et al., 2019 & Agarwal V et al., 2015). Furthermore, the hub mRNAs were obtained from the PPIN of the module-specific genes. The CC-specific dysregulated lncRNA-miRNA-mRNA network was constructed using these interaction partners. The significant molecular entities in the ceRNA network were identified by performing network analysis using the igraph R package and visualized with Cytoscape.

5.1.6 Functional enrichment analysis

Key genes and lncRNAs identified from regulatory networks were assessed for their functions through gene ontology terms and KEGG pathways. The clusterProfiler R package was used to conduct functional enrichment analysis. A p-value of 0.05 was the basis for filtering significant GO terms and KEGG pathways (Wu et al., 2021).

5.2 Results

5.2.1 Differentially expressed protein-coding genes and non-coding RNAs (miRNA & lncRNA)

The differential expression analysis on the expression profiles of protein-coding genes (mRNA) and non-coding RNA (miRNA & lncRNA) in cervical cancer tumor samples and normal samples was performed using the processed gene expression dataset. Significantly differentially expressed (DE) genes were identified using criteria of absolute log2 fold change > 2 and adjusted p-value < 0.01. 3661 differentially expressed mRNAs (DEmRNAs) in tumor samples were identified, including 1801 upregulated and 1860 downregulated. Similarly, 851 differentially expressed lncRNAs (DElncRNAs) were found, comprising 542 upregulated and 309 downregulated. In addition, 172 differentially expressed miRNAs (DEmiRNAs) were detected, consisting of 103 upregulated and 69 downregulated [Figure 1].

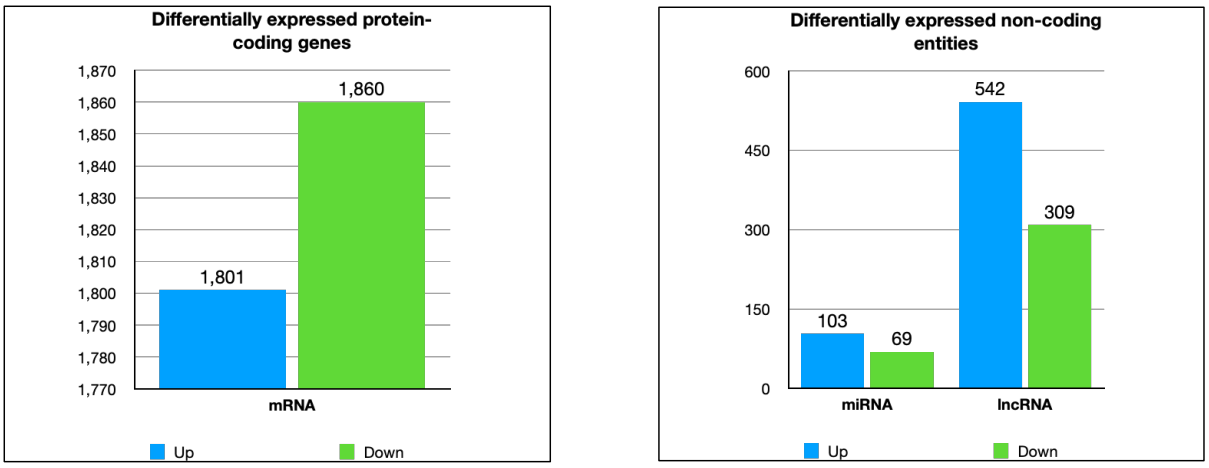


Figure 1: Differential expression analysis: Number of genes identified as differentially expressed among protein-coding genes and non-coding entities.

5.2.2 Identification of co-expressed modules from gene co-expression network

Weighted gene co-expression network analysis (WGCNA) was used to build a network of genes based on their pairwise correlations in order to elucidate the co-expression patterns among the genes from the differential mRNA expression data. The scale-free topology criterion was satisfied by a soft threshold power of 6, and 25 gene modules with distinct co-expression profiles were obtained [Figure 2]. A unique colour was assigned to each module, and the grey module stored genes that did not have co-expression patterns. Modules with similar expression patterns were merged by performing dynamic tree-cut analysis. Eigengene adjacency heatmap was generated by taking all the pairwise correlations among the modules [Figure 3]. The number of genes in each module was shown in barplot [Figure 4].

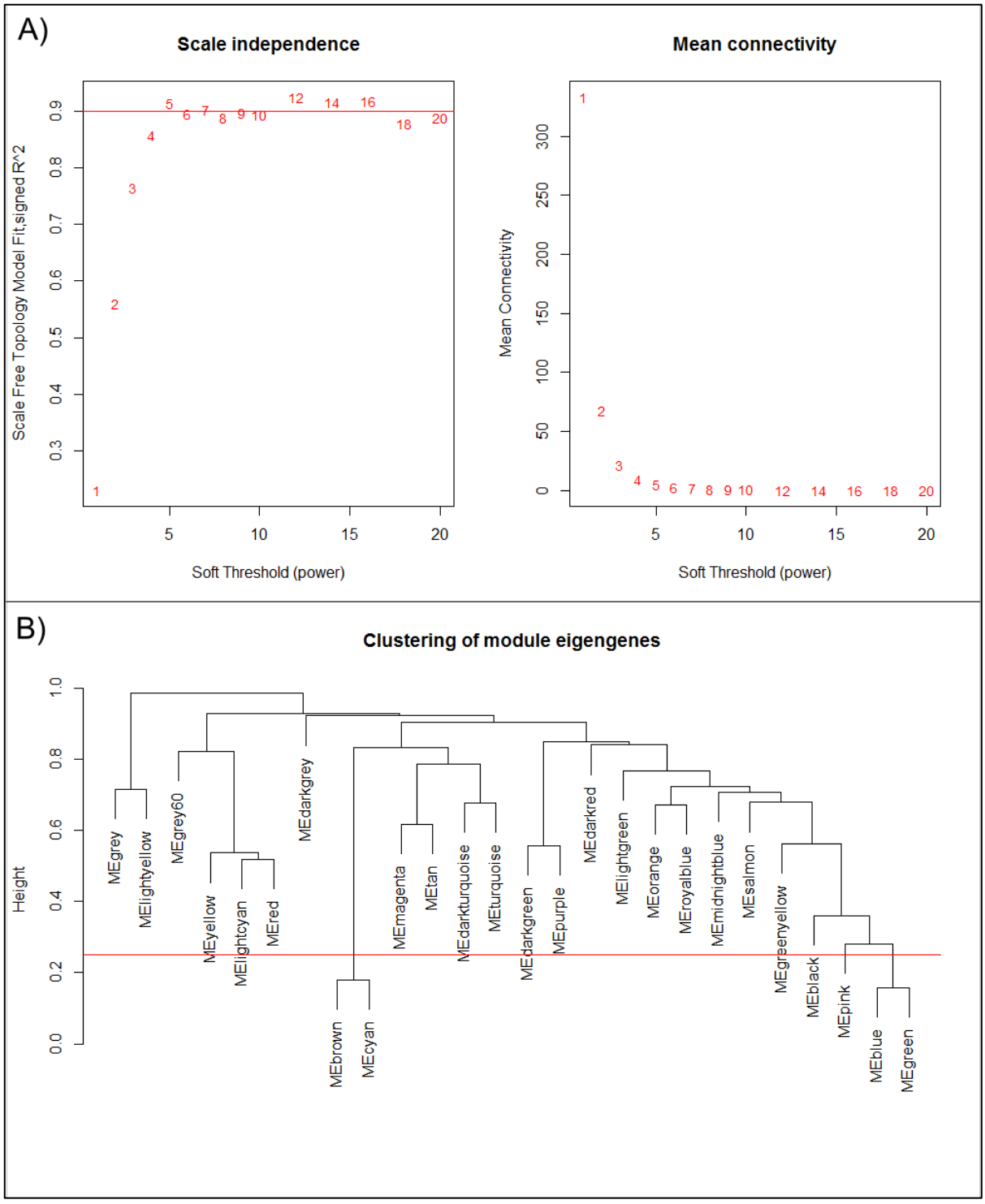


Figure 2: Weighted gene co-expression network construction: A) Scale-free topology plot for all the pairwise correlations B) Hierarchical clustering of co-expressed modules

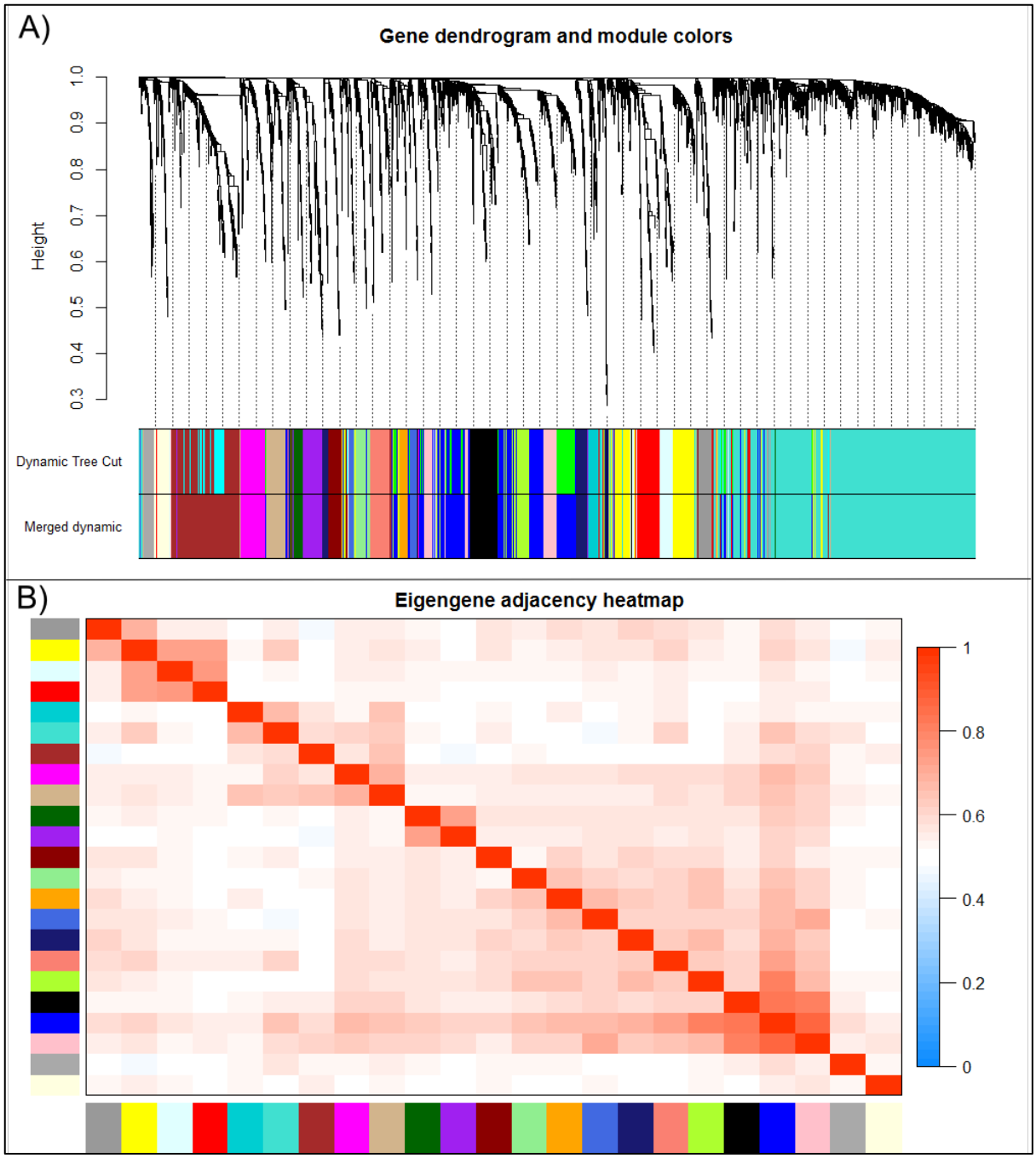


Figure 3: A) Cluster dendrogram with module colours B) Eigengene adjacency heatmap of all the pairwise correlations among the modules

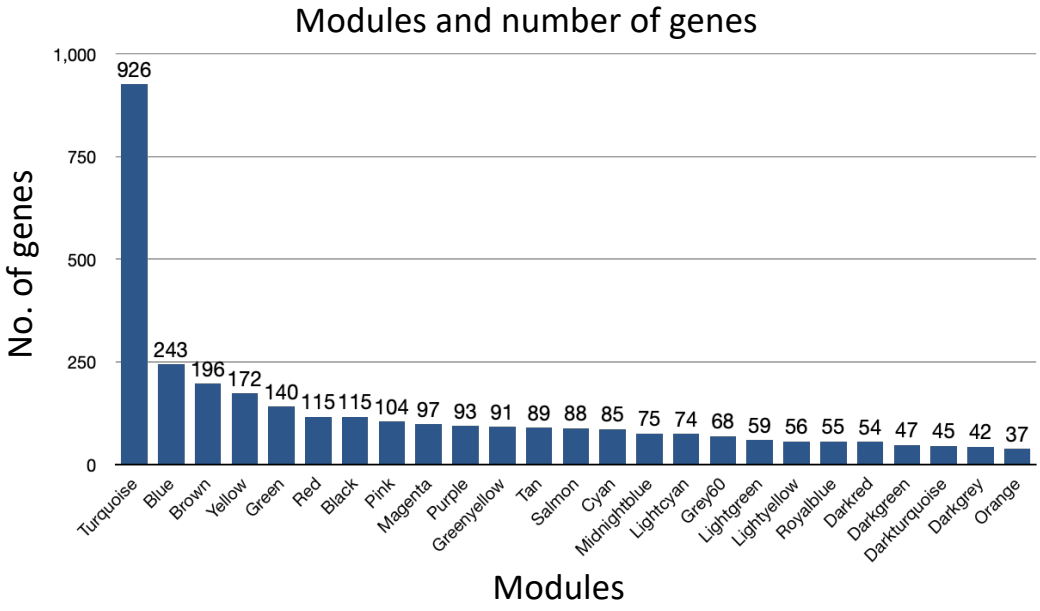


Figure 4: detected modules based on dynamic tree cut algorithm with no.of genes with co-expression patterns

5.2.3 Protein-protein interaction network analysis

Cervical cancer progression-related PPI network consists of 2487 nodes and 11574 interactions, that is extracted from the core human PPI network consisting of 51209 nodes with 8,52,432 edges [Figure 5a]. Network topological structural analysis of the PPI network results in the identification of 151 proteins as hubs and 131 as bottlenecks. It was observed that 92 proteins possessed both hub and bottleneck properties [Figure 5b]. Hubs and bottlenecks are important in protein-protein interaction networks because they are the proteins that have the most interactions with other proteins. Hubs are proteins that have many interactions with other proteins, while bottlenecks are proteins that connect different parts of the network together. In diseases, these hubs and bottlenecks can be targeted to help treat the disease.

For the module-specific PPI network, we found three significant modules: turquoise (884 nodes and 2312 edges), blue (138 nodes and 128 edges), and brown (44 nodes and 70 edges). These modules may represent distinct biological processes or pathways related to the DEmRNAs.

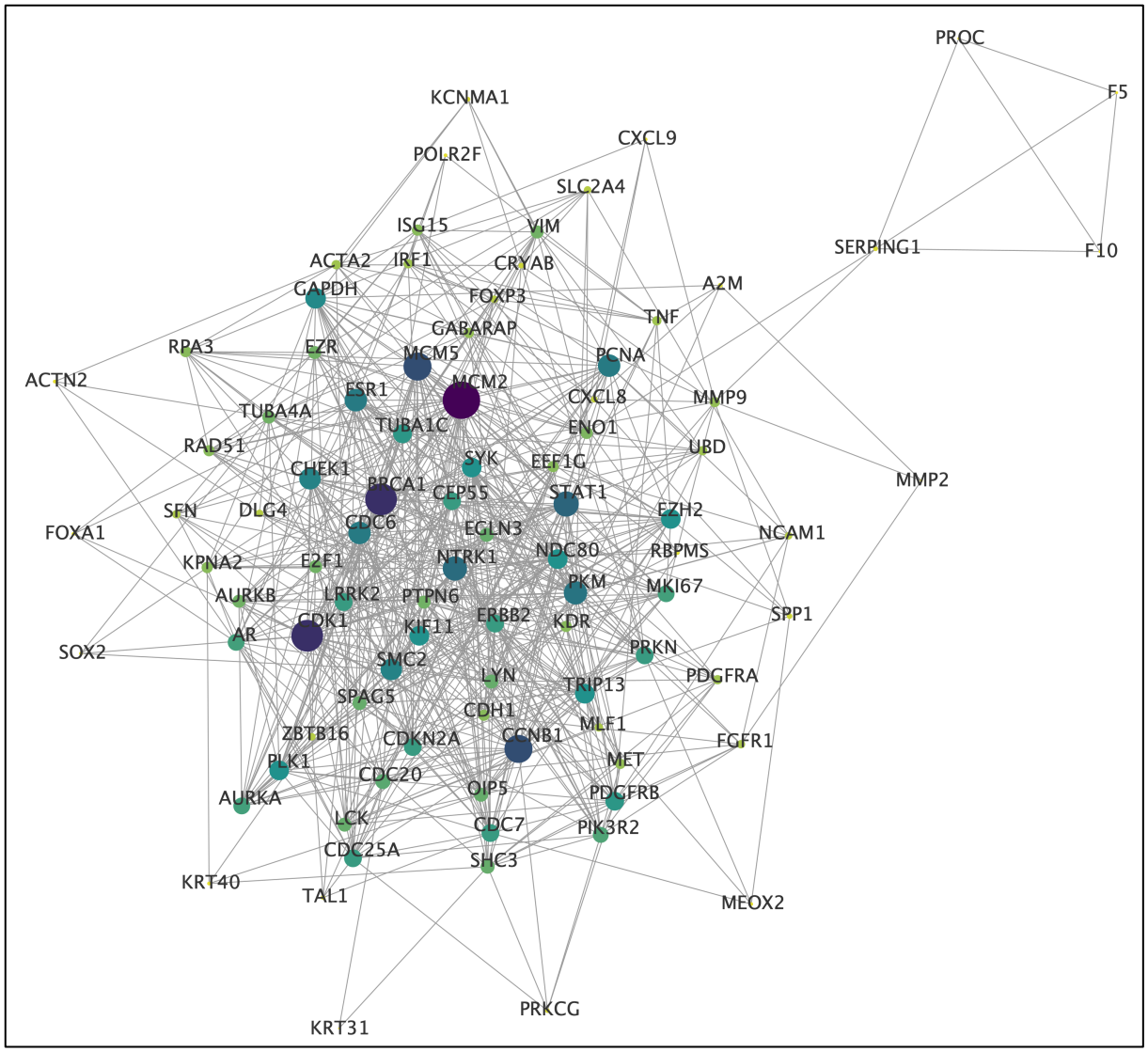


Figure 5a: Dysregulated PPI network with hubs and bottlenecks represented with varying node size

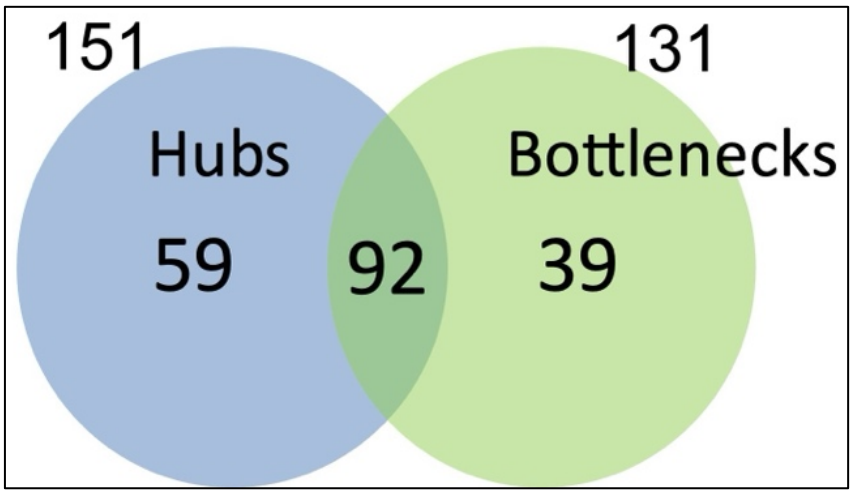


Figure 5b: Venn diagram illustrating the no. of hubs and bottlenecks.

5.2.3.1 Topological structural properties of the PPI network

To identify/assess the proteins that influence overall network structural stability, topological structural properties such as degree, betweenness and clustering coefficient centralities were calculated. The PPI network's degree distribution is scale-free and follows the power law. It indicates that higher connections are held by a small number of nodes in the network and less connections are held by a higher number of nodes. These essential genes are functionally significant in multiple pathways and contribute to the network's resilience against external perturbations. Degree vs degree distribution, degree vs betweenness and degree vs clustering coefficient properties were compared to understand the biological network behavior. We observed that dysregulated network follows properties of biological networks, degree vs degree distribution follows power-law in nature, and degree vs betweenness are positively correlated, indicating that high degree nodes and nodes with high betweenness are critical for information flow in network and degree vs clustering coefficient are negatively correlated that shows that a higher clustering coefficient values may indicate the presence of functional modules or pathways [Figure 6;7].

Further, the cervical cancer gene database (CCDB) was used to investigate genes with experimental validation and not well documented for the disease progression. Out of 92 genes that were both hubs and bottlenecks, having higher connections with other genes and controlling information flow between the genes, 24 genes were well-established with the disease progression, and no reports were found for the remaining 68 genes. These genes were enriched in significant gene ontology terms of biological processes (1048 terms) and significant KEGG pathway terms (78 terms).

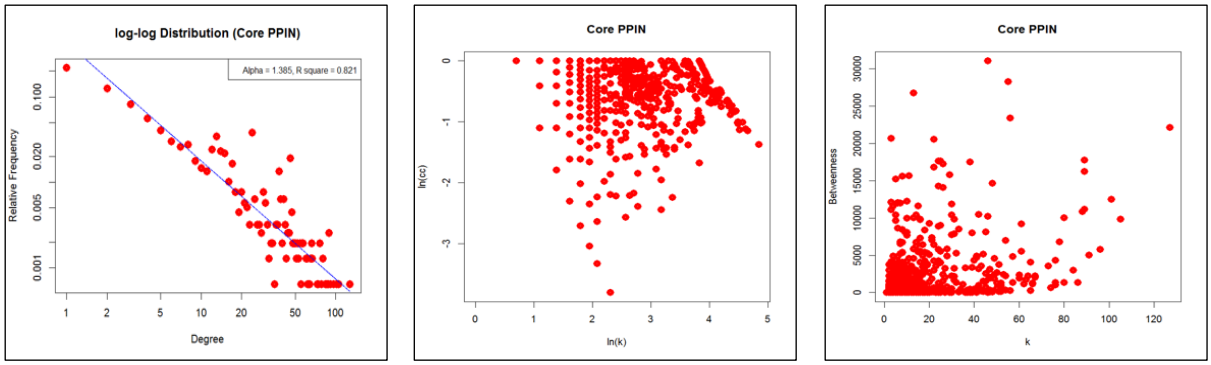


Figure 6: Topological properties of core PPI network: The X-axis indicates degree and the Y-axis indicates relative frequency for the left panel; clustering coefficient in the middle panel; betweenness in the right panel

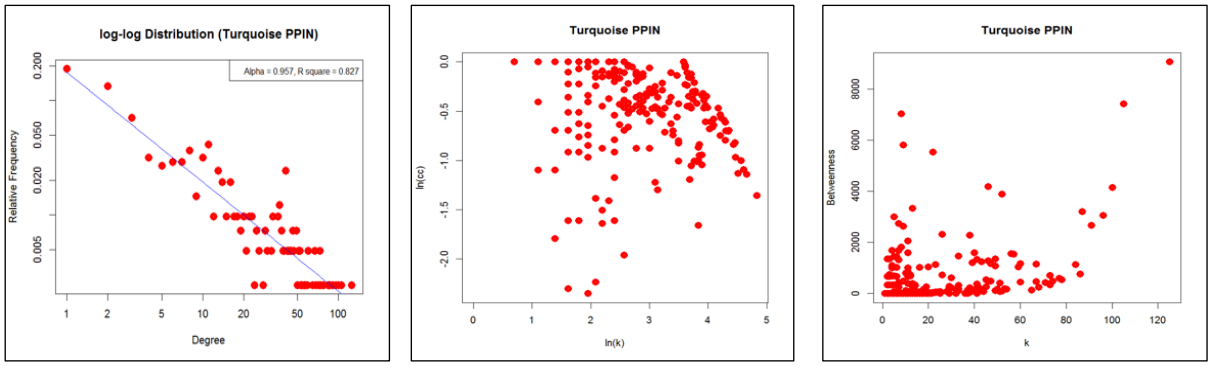


Figure 7: Topological properties of Turquoise module network: The X-axis indicates degree and the Y-axis indicates relative frequency for the left panel; clustering coefficient in the middle panel; betweenness in the right panel

5.2.4 Highly correlated co-expressed modules from Module – clinical feature associations

Module-clinical feature relationships analysis identified modules that observed a significant correlation with HPV status: turquoise (r=-0.34, p=1e-09), yellow (r=-0.18, p=0.002), brown (r=-0.18, p=0.002), blue (r=-0.15, p=0.007), lightcyan (r=-0.16, p=0.004), pink (r=-0.16, p=0.004), and lightyellow (r=-0.13, p=0.02) [Figure 8]. These modules may contain differentially expressed mRNAs that play critical roles in regulating cervical cancer tumorigenesis. GO terms and pathways were assessed for each module to elucidate their biological functions and pathways.

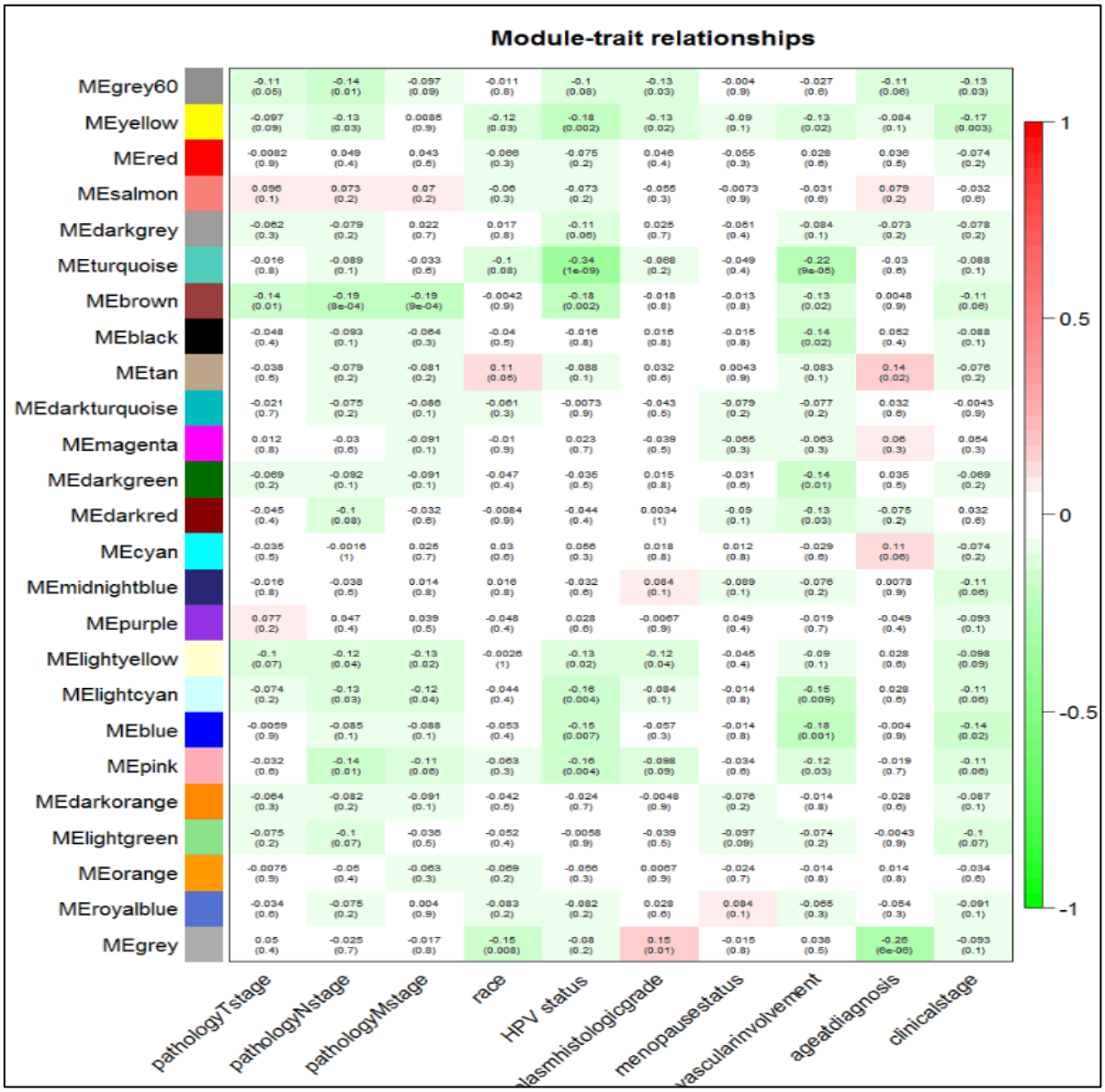


Figure 8: Module-trait associations

5.2.5 Integrative regulatory network analysis

Gene regulatory networks of direct/indirect and sponging mechanisms were analyzed to understand the transcriptional and post-transcriptional gene regulation in the disease state. In the below sections, the direct/indirect regulation of the lncRNA-mRNA co-expression network is explained in detail, along with a discussion on the ceRNA network of lncRNA-miRNA-mRNA via the miRNA-mediated sponging mechanism.

5.2.5.1 Integrative lncRNA-miRNA-mRNA ceRNA network analysis

The competing endogenous RNA (ceRNA) regulatory network is composed of 98 nodes and 101 edges, involving 29 lncRNAs, 3 miRNAs and 66 mRNAs. The network consists of three subnetworks, in which each miRNA plays a central role by interacting with both lncRNAs and mRNAs. The subnetworks are interconnected by 5 lncRNAs (KCNQ1OT1, DLEU1, SNHG14, LINC00111 and TMEM72-AS1) and one mRNA (FSCN1), which act as bridges for information transfer [Figure 9].

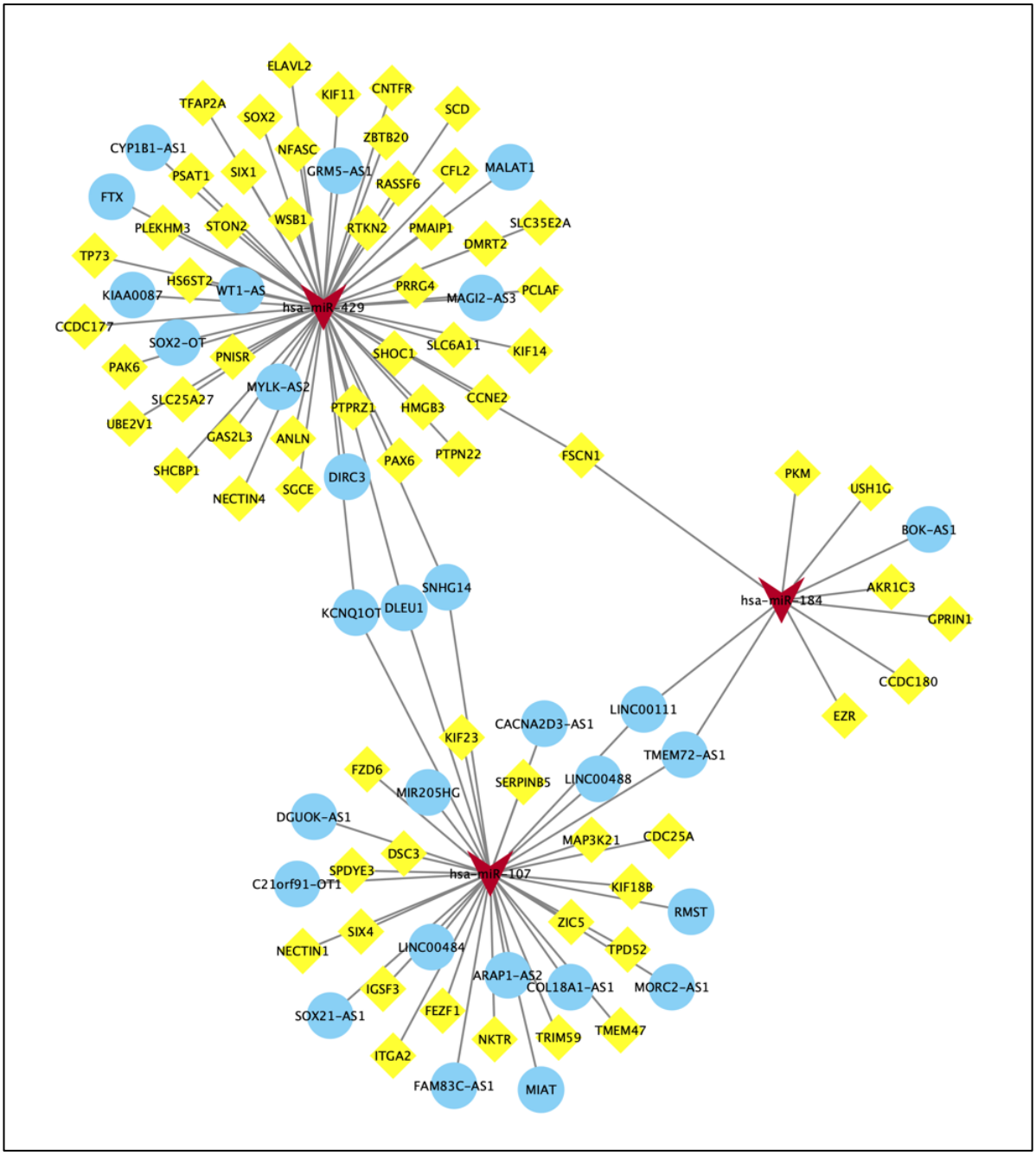


Figure 9: lncRNA-miRNA-mRNA ceRNA network of dysregulated coding and non-coding entities

To determine the key hubs that have critical functions in the network, various centrality measures were computed and found that lncRNAs KCNQ1OT1, SNHG14 and DLEU1 regulate several coding genes associated with different biological processes through miRNAs. The sponging mechanism refers to the process by which lncRNAs bind to miRNAs and prevent them from targeting mRNAs, thus modulating gene expression. In addition, DEmiRNAs hsa-miR-107, hsa-miR-184 and hsa-miR-429 play crucial roles in three subnetworks as central node.

5.2.5.2 Integrative lncRNA-mRNA regulatory network analysis

The lncRNA-mRNA regulatory network has 152 nodes and 206 regulatory interactions. Among the nodes, there are 50 lncRNAs and 102 mRNAs. The networks are directed, meaning that information flows from one node to another. The essential genes and lncRNAs, based on the number of connections they have were found by using degree centrality. Indegree is the incoming connections of a node, and outdegree is outgoing connections of a node. The genes with high indegree and outdegree are important for the network stability and function. These genes are EZH2, CDH1, BCL2, MMP9, ZEB1, MMP2 and VIM. They are targeted by many lncRNAs and are involved in key pathways. We also found the lncRNAs with high outdegree, which are MALAT1, CDKN2B-AS1, MEG3, HOTTIP, CYTOR, FEZF1-AS1 and FENDRR. Some of these lncRNAs (MALAT1, CDKN2B-AS1) have been linked to cervical cancer, while others have not been reported with the hub genes [Figure 10]. Interactions were further explored using the lncRNA-mRNA interaction database, revealing potential interaction site and binding energy values [Figure 11].

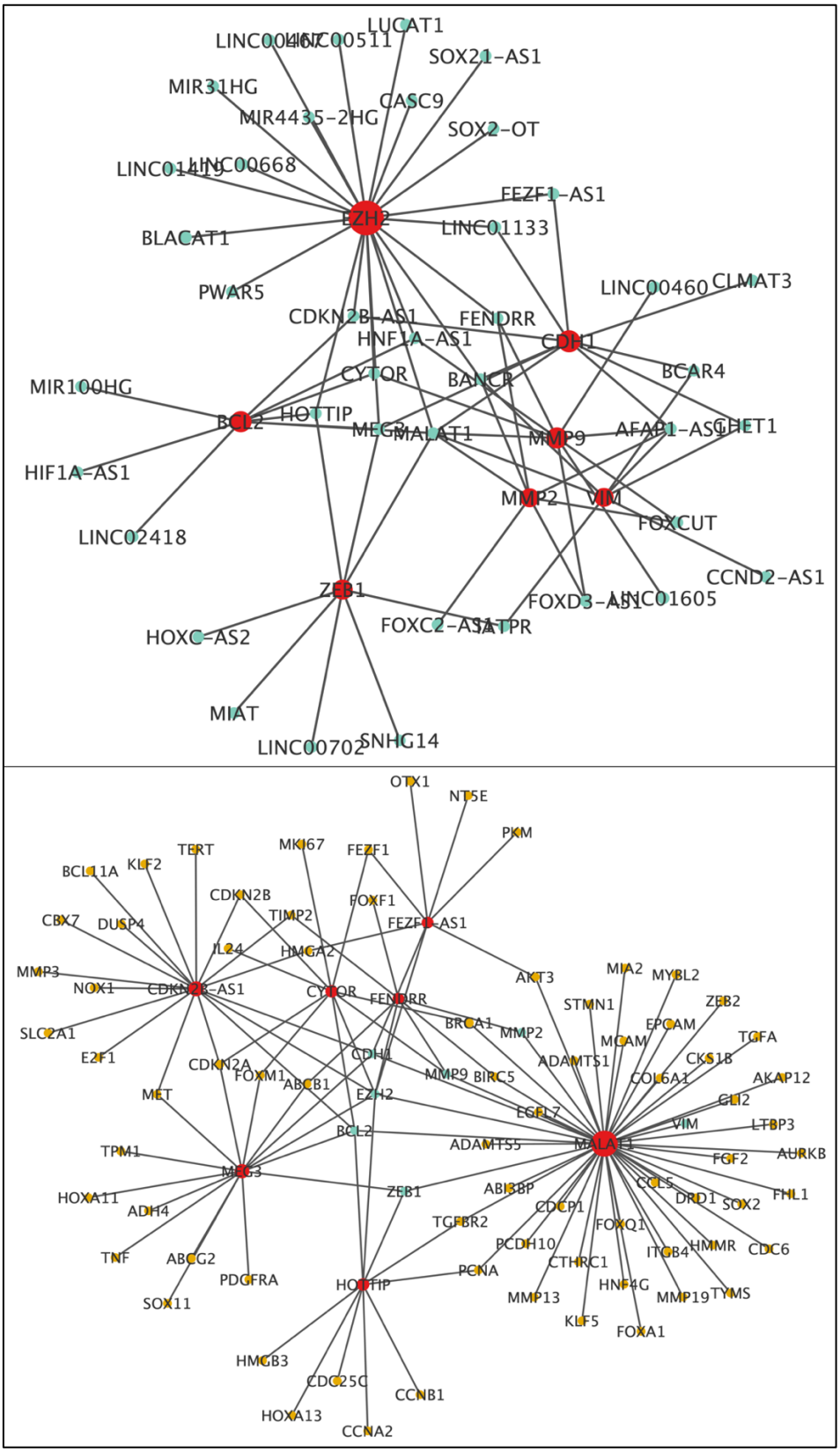


Figure 10: lncRNA-mRNA regulatory network: the top panel shows inward interaction towards protein-coding genes; the bottom panel shows outward interaction going from lncRNAs

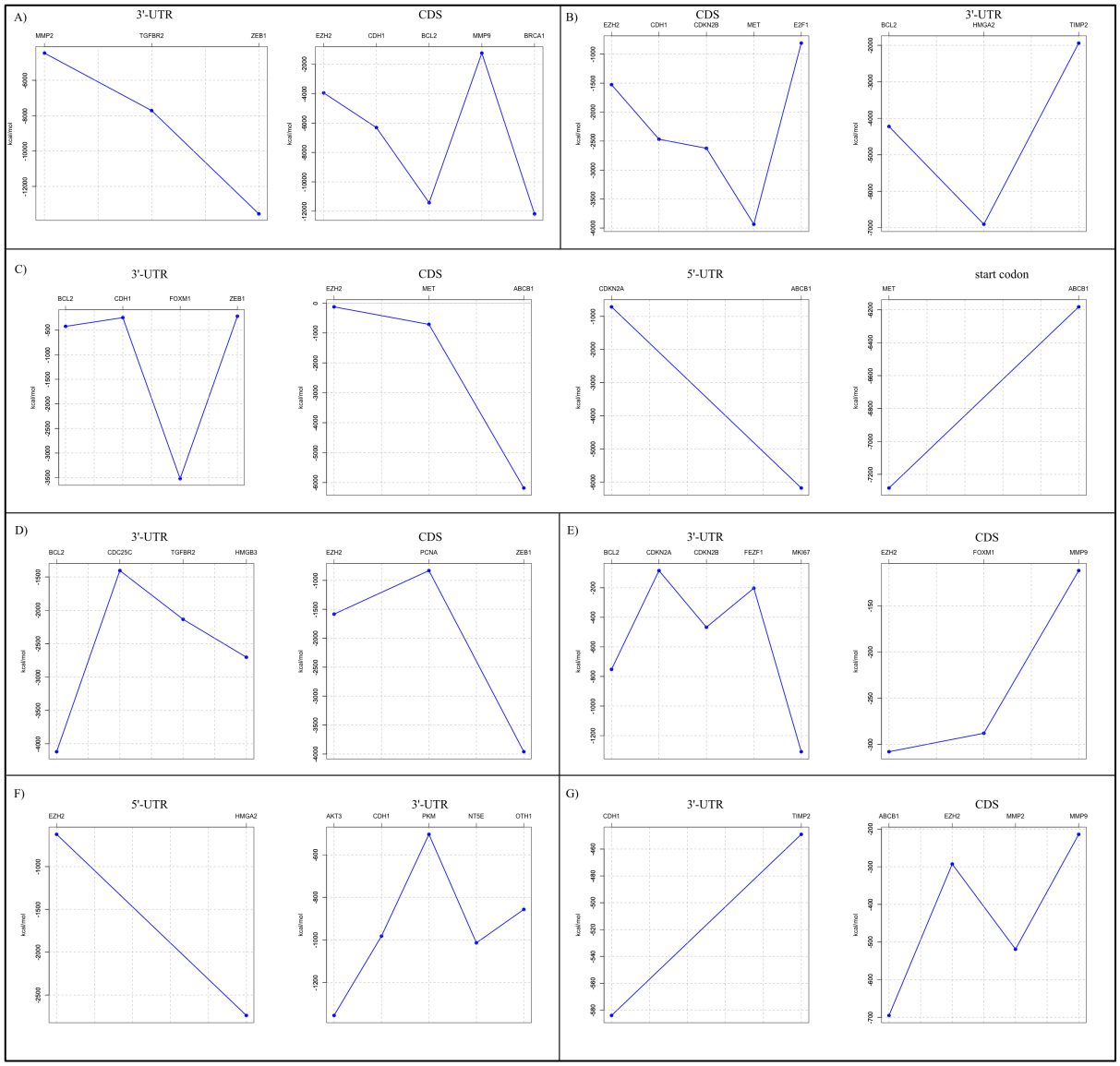


Figure 11: lncRNA-mRNA interaction and their binding sites: The binding energies of the corresponding complexes (sumenergy) indicate the strength of the interactions, with more negative binding energy implying stronger association and more favorable or stable interaction. A)MALAT1 B)CDKN2B-AS1 C)MEG3 D)HOTTIP E)CYTOR/LINC00152 F)FEZF1-AS1 G)FENDRR

5.2.6 Functional enrichment analysis

GO term and KEGG pathway analyses revealed that these genes participate in various biological processes such as regulating the G2/M phase transition of the cell cycle, positively regulating cell cycle processes, regulating nuclear division, cell cycle checkpoint, and mitotic

nuclear division. Moreover, several KEGG pathways related to these genes were also identified, such as cell cycle, Ras signaling pathway, viral carcinogenesis, proteoglycans in cancer, cellular senescence, glioma and prostate cancer [Figure 12]. These findings suggested that the key genes might play critical roles in the cell cycle regulation and the development of various cancers. Therefore, these genes might serve as potential biomarkers or therapeutic targets for cancer diagnosis and treatment.

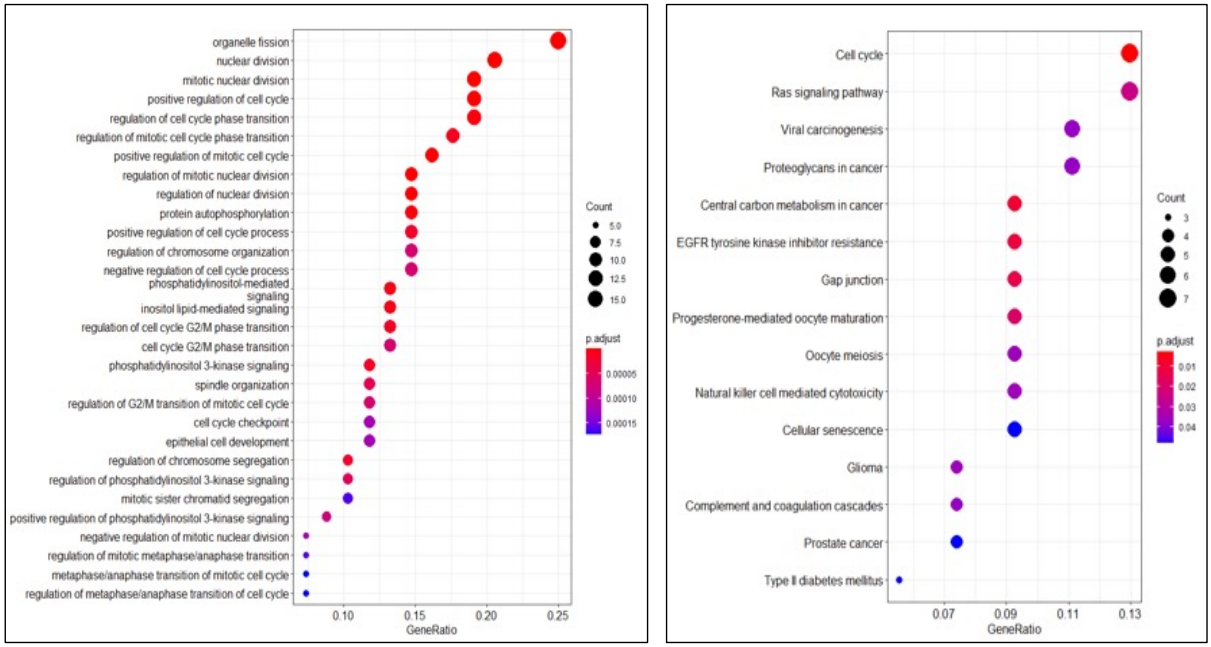


Figure 12: Gene Ontology terms and KEGG pathway analysis

5.3 Discussion

Cervical cancer is a common and deadly gynecological malignancy that affects women worldwide. One of the emerging mechanisms that regulates cervical cancer progression is the involvement of ncRNAs, such as miRNAs, lncRNAs, and circRNAs. These ncRNAs can modulate various cellular processes such as apoptosis, cell cycle, angiogenesis, invasion, and metastasis, either directly or indirectly. In this study, a network-based approach was applied to explore the key regulatory interactions and genes that influence tumor initiation and progression. Genes were grouped into modules by WGCNA method on the basis of their expression patterns and correlation with clinical traits. 6 co-expressed modules that were highly expressed and related to clinical traits were found. The turquoise module had the high number of genes. The PPI network analysis confirmed that the networks followed the scale-free

property of biological networks. Hubs and bottlenecks were also identified from the PPI networks and the target-gene interaction network.

The regulatory network analysis based on centrality measures revealed several genes as hubs, which also had regulatory interactions with ncRNAs, namely miRNA and lncRNA. Among these, some key mRNAs were EZH2, CDH1, BCL2, MMP9, ZEB1, MMP2, and VIM. These mRNAs are associated with various biological processes such as apoptosis, cell proliferation, invasion, and metastasis. Some important lncRNAs were MALAT1, CDKN2B-AS1, MEG3, HOTTIP, CYTOR, FEZF1-AS1, and FENDRR. These lncRNAs act as regulators of gene expression by modulating the chromatin structure, transcriptional machinery, or post-transcriptional events. Further, we identified three miRNAs, hsa-miR-107, hsa-miR-184 and hsa-miR-429, that are involved in the regulation of key mRNA and lncRNA molecules. These were the main components of the lncRNA-mRNA interaction network and lncRNA-miRNA-mRNA ceRNA regulatory network, which could provide better understanding of the molecular mechanisms of cervical cancer development and progression.

EZH2, a histone methyltransferase, has a vital role in tumor progression by promoting cell survival, proliferation, and invasion (Gan et al., 2018). It is upregulated in various cancer types, making it a potential target for anticancer therapy (Shen et al., 2013). CDH1, a tumor suppressor gene, is frequently hypermethylated in multiple cancers such as breast (Huang et al., 2015), head and neck (Shen et al., 2016), and esophageal (Ling et al., 2011). BCL-2 is implicated in the progression of various cancers, including prostate, breast, and chronic lymphocytic leukemia (Adams and Cory, 2018). Its overexpression in cancer cells inhibits apoptosis, promoting their survival and growth (Radha and Raghavan 2017). MMP2 and MMP9, belongs to the matrix metalloproteinase (MMP) family, has a key role in cancer progression, particularly in angiogenesis, tumor growth, and metastasis (Klein et al., 2004). In colorectal cancer, MMP2 and MMP9 are involved in epithelial-to-mesenchymal transition and immune response, suggesting their potential as biomarkers (Buttacavoli et al., 2021). ZEB1 is a key regulator of epithelial-to-mesenchymal transition (EMT), which is a process that contributes to cancer progression and metastasis (Caramel et al., 2018). ZEB1 influences the expression of genes related to EMT, stem cell properties, immune escape, and epigenetic modifications. ZEB1 also contributes to the silencing of E-cadherin, a tumor suppressor gene,

through its interaction with chromatin-modifying enzymes (Zhang et al., 2019). Vimentin, an intermediate filament protein, plays a crucial role in tumor progression, particularly in tumor growth promotion, invasion, and metastasis (Satelli and Li 2011). This is further supported by the finding that vimentin is overexpressed in various cancers, including colorectal cancer, where it induces tumor growth and metastasis via epithelial-to-mesenchymal transition (EMT) (Strouhalova et al., 2018)

MALAT1 is a lncRNA associated with various malignancies (Hao et al., 2023). It is upregulated in lung, breast and colorectal cancers. It promotes cancer cell proliferation, migration and invasion. It is related to the poor prognosis and promotes cancer cell migration and metastasis by inducing epithelial-mesenchymal transition (EMT) in lung cancer (Shen et al., 2015). It has been implicated in breast cancer and has been the subject of diagnostic and prognostic studies (Jiang et al., 2020). It is closely related to the cell proliferation, tumorigenicity, and metastasis in colorectal cancer (CRC). It targets various signaling pathways and microRNAs, playing a pivotal role in CRC pathogenesis (Xu et al., 2022). It has been shown to impact the differentiation of effector and memory CD8+ T cell subsets by mediating epigenetic repression of memory-associated genes in terminal effector cells (Kanbar et al., 2022). MALAT1 has been found to modulate Smad1, contributing to colorectal cancer progression by regulating autophagy (Zhou et al., 2021). CDKN2B-AS1 was found to be associated with atherosclerosis, diabetes, and alzheimer's disease. It has been found to be downregulated in glioma and implicated in the disease's progression (Bi et al., 2018).

MEG3 has been found to be downregulated in hepatocellular carcinoma, glioma, and ovarian cancer and acts as a tumor suppressor by inhibiting cancer cell proliferation and inducing apoptosis (Xu et al., 2022). HOTTIP is upregulated in colorectal, pancreatic and ovarian cancers and promotes cancer cell proliferation, migration, and invasion (Liu et al., 2020). CYTOR has been found to be upregulated in breast, lung and colorectal cancers and promotes cancer cell proliferation and migration (Tian et al., 2021). FEZF1-AS1 has been associated with glioma, breast cancer, and colorectal cancers and promotes cell proliferation and migration (Zhou et al., 2019). FENDRR is associated with lung, colorectal, and ovarian cancers. It has been shown to act as a tumor suppressor by inhibiting cell proliferation and inducing apoptosis (Zheng et al., 2021; Jiang et al., 2020).

miR-107 acts as an oncogene, promoting the growth and metastasis of gastric cancer by downregulating FAT4 and activating the PI3K/AKT signaling pathway. On the other hand, It can also act as a tumor suppressor, inhibiting the proliferation and invasion of prostate cancer by targeting CDC42 and suppressing the Rho GTPase signaling pathway (Chen et al., 2021; Fan et al., 2020). miR-184 has been reported by several studies that it acts as a tumor suppressor in various types of cancer, such as nasopharyngeal carcinoma, colorectal cancer, and lung adenocarcinoma, by targeting oncogenic factors or signaling pathways involved in tumorigenesis (Wu et al., 2017; Rao et al., 2022). miR-429 plays a crucial role in maintaining epithelial phenotype and preventing epithelial-mesenchymal transition (EMT), facilitating tumor invasion and metastasis. It is reported to act as a tumor suppressor or an oncogene in various cancers, including endometrial, gastric, ovarian and colorectal cancers (Leet et al., 2023). Some important genes/lncRNAs might not be detected due to constraints such as the p-value threshold applied and the proposed hypotheses here need further verification by experimental methods to better understand their role in tumor progression.

5.4 Conclusion

This work explores an integrative network approach to understand the regulatory interactions between ncRNAs and mRNAs in cervical cancer progression. After analyzing their interaction patterns, multiple key genes and ncRNAs (miRNA & lncRNA) were identified as master regulators that may play crucial roles in the gene regulation network of this disease and may offer novel drug targets for therapeutic intervention. The key lncRNAs also have potential as prognostic markers of cancer outcomes and as predictive and diagnostic tools for cancer detection.

Chapter 6

Summary and Conclusion

High-throughput sequence technologies have revolutionized the discovery of cancer-driver genes. Unlike the traditional method of characterizing individual genes, which was slow and laborious, these technologies can examine many genes simultaneously and reveal novel ones involved in cancer progression. These technologies provide a comprehensive and rapid insight into the complex biological process in cancer cells.

Most cancer-driver gene prediction or identification focuses on protein-coding genes, which make up 2% of the human genome. However, recent studies reveal that genetic alterations also affect non-coding regions comprising 98% of the genome. These findings suggest that non-coding RNAs, along with protein-coding genes, can influence tumor development and progression. Yet, only a few ncRNAs, especially lncRNAs, have been identified and characterized in multiple cancers. These can regulate gene expression at multiple levels, epigenetic modifications, signaling and metabolic pathways.

This study aimed to prioritize potential candidate genes and identify lncRNA as a master gene regulator for cervical squamous cell carcinoma. To achieve this, composition-based and k-mer frequency-based alignment-free sequence analyses were conducted to propose candidate genes for cervical cancer. Additionally, a comprehensive analysis of differentially expressed genes and ncRNAs and their interactions were analyzed using network analysis to understand the molecular mechanisms at the systems level. The interactions between these genes and non-coding RNAs were explored through co-expression and protein-protein interaction analysis, and target-gene interaction network analysis was employed to elucidate the regulatory mechanisms and pathways involved.

Our first two objectives deal with the prioritization of candidate genes associated with cervical cancer by studying the sequence profile of gene and protein sequences. Alignment-free methods employed to analyze sequence similarity between the cancer driver genes and candidate cancer genes. The first objective deals with the analysis of sequence similarity among both sets of protein sequences based on the amino acid composition, i.e., physicochemical properties of amino acids. 14 potential candidate genes with high similarity scores with cancer driver genes were identified that might be considered for further experimental validation. Gene ontology analysis revealed their significance in cell cycles and regulation of granulocyte differentiation, either positively regulating or negatively in several cell types that include

epithelial cells, muscle cells, leukocytes, and lymphocytes. KEGG pathway analysis reveals these genes involved acute myeloid leukemia, chronic myeloid leukemia, transcriptional misregulation in cancer, prostate cancer, and FoxO signaling pathway.

The second objective deals with fractal analysis of the cancer driver genes and candidate cancer genes to prioritize potential candidates for cervical cancer based on their correlation. It results in the identification of 16 prioritized genes that have a high correlation with known cancer genes. It was also observed that both approaches prioritized four genes. These prioritized genes were associated with transcriptional misregulation in cancer, PI3K-Akt signaling pathway, ErbB signaling pathway and cell cycle-related GO terms.

Our study's third and fourth objectives aimed to construct and analyze gene expression-based interaction networks to elucidate the key molecular players, including protein-coding and non-coding elements, associated with the progression of cervical cancer. To achieve the third objective, we analyzed three gene expression datasets of cervical cancer to reconstruct and analyze the protein-protein interaction network of the differentially expressed genes. Network centrality measures, such as hub-bottleneck and relative vulnerability analyses, were applied to identify these key genes. These genes are MCM5, FN1, TRIP13, KIF11, TTK, CDC45, and BUB1B. We further validated the functional relevance of these genes by investigating their association with important biological processes and pathways. Moreover, these genes expression at both mRNA and protein levels were confirmed using immunohistochemistry data. Higher expression levels of these proteins were observed to be significantly associated with poor prognosis of cervical cancer patients. These proteins could serve as candidate biomarkers for survival prediction.

The fourth objective aimed to elucidate how ncRNAs, especially long ncRNAs (lncRNAs), modulate gene expression in various biological processes. To achieve this, gene co-expression networks and ncRNA-gene interaction networks were constructed. The centrality analysis was performed to identify the key nodes and edges that influence the network topology and functionality. Several lncRNAs with high outdegree centrality, such as MALAT1, CDKN2B-AS1, MEG3, HOTTIP, CYTOR, FEZF1-AS1 and FENDRR, were identified, indicating that multiple target genes were regulated by them through diverse mechanisms. Conversely, it was found that some genes, such as EZH2, CDH1, BCL2, MMP9, ZEB1, MMP2 and VIM, had

high in-degree centrality, implying that they were involved in critical cellular processes and were regulated by numerous lncRNAs. Furthermore, the binding sites and energy of the lncRNA-mRNA interactions were analyzed and it was observed that most of the lncRNAs interacted with the 3' untranslated region (3'UTR) of their target genes with low binding energy, suggesting a post-transcriptional mechanism of regulation.

This study provides a detailed understanding of candidate gene prioritization and key molecular players identification by exploring differential gene expression patterns and centrality-based network analysis that, in turn, leads to lncRNA-mediated regulation in cancer. The study augments the comprehension of lncRNA's role in cancer etiology and proliferation.

Bibliography

Aalijahan, H., & Ghorbian, S. (2019). Long non-coding RNAs and cervical cancer. Experimental and molecular pathology, 106, 7-16.

Adams, J.M., & Cory, S. (2018). The BCL-2 arbiters of apoptosis and their growing role as cancer targets. Cell death and differentiation, 25 1, 27-36 .

Adie EA, Adams RR, Evans KL, Porteous DJ, Pickard BS. SUSPECTS: enabling fast and effective prioritization of positional candidates. Bio- informatics. 2006;22:773-774. https://doi.org/10.1093/ bioinformatics/btk031

Aerts S, Lambrechts D, Maity S, et al. Gene prioritization through genomic data fusion. Nat Biotechnol. 2006;24:537-544. https://doi. org/10.1038/nbt1203

Agarwal SM, Raghav D, Singh H, Raghava GPS. CCDB: A curated database of genes involved in cervix cancer. Nucleic Acids Res. 2011;39(SUPPL. 1):975-979. doi:10.1093/nar/gkq1024

Agarwal V, Bell GW, Nam J, Bartel DP. Predicting effective microRNA target sites in mammalian mRNAs. eLife, 4:e05005, (2015). eLife Lens view.

Ala, U. (2020). Competing endogenous RNAs, non-coding RNAs and diseases: an intertwined story. Cells, 9(7), 1574.

Alonso-López D, Gutiérrez MA, Lopes KP, Prieto C, Santamaría R, De Las Rivas J. APID interactomes: Providing proteome-based interactomes with controlled quality for multiple species and derived networks. Nucleic Acids Res. 2016;44(W1). doi:10.1093/nar/gkw363

Alonso-López Di, Campos-Laborie FJ, Gutiérrez MA, et al. APID database: Redefining protein-protein interaction experimental evidences and binary interactomes. Database. 2019;2019. doi:10.1093/database/baz005

Amala, A., & Emerson, I. A. (2019). Identification of target genes in cancer diseases using protein–protein interaction networks. Network Modeling Analysis in Health Informatics and Bioinformatics, 8, 1-13. https://doi.org/10.1007/s13721-018-0181-1

Araujo AM, Abaurrea A, Azcoaga P, et al. Stromal oncostatin M cytokine promotes breast cancer progression by reprogramming the tumor microenvironment. J Clin Invest. 2022;132(7). doi:10.1172/JCI148667

Arneodo, A., Grasseau, G., & Holschneider, M. (1988). Wavelet transform of multifractals. Physical review letters, 61(20), 2281.

Artibani M, Sims AH, Slight J, et al. WT1 expression in breast cancer disrupts the epithelial/mesenchymal balance of tumour cells and cor- relates with the metabolic response to docetaxel. Sci Rep. 2017;7: 45255. https://doi.org/10.1038/srep45255

Ashouri, A., Sayin, V., Van den Eynden, J. et al. Pan-cancer transcriptomic analysis associates long non-coding RNAs with key mutational driver events. Nat Commun 7, 13197 (2016). https://doi.org/10.1038/ncomms13197

Azadifar, S., Ahmadi, A. A novel candidate disease gene prioritization method using deep graph convolutional networks and semi-supervised learning. BMC Bioinformatics 23, 422 (2022). https://doi.org/10.1186/s12859-022-04954-x

Bai F, Wang T. On graphical and numerical representation of protein sequences. J Biomol Struct Dyn. 2006;23:537-545. https://doi.org/10. 1080/07391102.2006.10507078

Baltimore, D. Our genome unveiled. Nature 409, 815–816 (2001). https://doi.org/10.1038/35057267

Barabási AL, Gulbahce N, Loscalzo J. Network medicine: A network-based approach to human disease. Nat Rev Genet. 2011;12(1). doi:10.1038/nrg2918

Barabási AL, Oltvai ZN. Network biology: understanding the cell's functional organization. Nat Rev Genet. 2004 Feb;5(2):101-13. doi: 10.1038/nrg1272.

Bartel D. P. (2004). MicroRNAs: genomics, biogenesis, mechanism, and function. Cell, 116(2), 281–297. https://doi.org/10.1016/s0092-8674(04)00045-5

Batista, P. J., & Chang, H. Y. (2013). Long noncoding RNAs: cellular address codes in development and disease. Cell, 152(6), 1298-1307.

Bazzini, A. A., Johnstone, T. G., Christiano, R., Mackowiak, S. D., Obermayer, B., Fleming, E. S., ... & Giraldez, A. J. (2014). Identification of small ORF s in vertebrates using ribosome footprinting and evolutionary conservation. The EMBO journal, 33(9), 981-993.

Benzi G, Camasses A, Atsunori Y, Katou Y, Shirahige K, Piatti S. A common molecular mechanism underlies the role of Mps1 in chromosome biorientation and the spindle assembly checkpoint. EMBO Rep. 2020;21(6). doi:10.15252/embr.202050257

Berti, F. C. B., Lobo-Alves, S. C., Oliveira-Tore, C. D. F., Salviano-Silva, A., de Oliveira, K. B., de Araujo-Souza, P. S., ... & Malheiros, D. (2021). Competing Endogenous RNAs in Cervical Carcinogenesis: A New Layer of Complexity. Processes, 9(6), 991.

Bi, Y-Y, Shen, G, Quan, Y, Jiang, W, Xu, F. Long noncoding RNA FAM83H-AS1 exerts an oncogenic role in glioma through epigenetically silencing CDKN1A (p21). J Cell Physiol. 2018; 233: 8896–8907. https://doi.org/10.1002/jcp.26813

Bray F, Ferlay J, Soerjomataram I, Siegel RL, Torre LA, Jemal A. Global cancer statistics 2018: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. CA Cancer J Clin. 2018;68(6):394-424. doi:10.3322/caac.21492

Burk RD, Chen Z, Saller C, et al. Integrated genomic and molecular characterization of cervical cancer. Nature. 2017;543:378-384. https://doi.org/10.1038/nature21386

Bustin SA, Dorudi S. Gene expression profiling for molecular staging and prognosis prediction in colorectal cancer. Expert Rev Mol Diagn. 2004;4(5). doi:10.1586/14737159.4.5.599

Buttacavoli, M., Di Cara, G., Roz, E., Pucci-Minafra, I., Feo, S., & Cancemi, P. (2021). Integrated multi-omics investigations of metalloproteinases in colon cancer: Focus on MMP2 and MMP9. International journal of molecular sciences, 22(22), 12389.

Cao X, Tang Z, Huang F, Jin Q, Zhou X, Shi J. High TMPRSS11D protein expression predicts poor overall survival in non-small cell lung cancer. Oncotarget. 2017;8(8). doi:10.18632/oncotarget.14559

Caramel, J., Ligier, M., & Puisieux, A. (2018). Pleiotropic roles for ZEB1 in cancer. Cancer research, 78(1), 30-35.

Carter SL, Eklund AC, Kohane IS, Harris LN, Szallasi Z. A signature of chromosomal instability inferred from gene expression profiles predicts clinical outcome in multiple human cancers. Nat Genet. 2006;38(9):1043-1048. doi:10.1038/ng1861

Chen M, Pan H, Sun L, et al. Structure and regulation of human epithelial cell transforming 2 protein. Proc Natl Acad Sci U S A. 2020;117(2). doi:10.1073/pnas.1913054117

Chen SJ, Liao DL, Chen CH, Wang TY, Chen KC. Construction and Analysis of Protein-Protein Interaction Network of Heroin Use Disorder. Sci Rep. 2019;9(1). doi:10.1038/s41598-019-41552-z

Chen X, Xiang H, Yu S, Lu Y, Wu T. Research progress in the role and mechanism of Cadherin-11 in different diseases. J Cancer. 2021;12(4). doi:10.7150/JCA.52720

Chen, L. Q., Yi, C. L., Liu, D. C., Wang, P., Zhu, Y. F., & Yuan, L. P. (2021). Hsa_circ_0041103 induces proliferation, migration and invasion in bladder cancer via the miR-107/FOXK1 axis. European review for medical and pharmacological sciences, 25(3), 1282–1290. https://doi.org/10.26355/eurrev_202102_24832

Chen, W., Chen, X., Wang, Y., Liu, T., Liang, Y., Xiao, Y., & Chen, L. (2019). Construction and Analysis of lncRNA-Mediated ceRNA Network in Cervical Squamous Cell Carcinoma by Weighted Gene Co-Expression Network Analysis. Medical science monitor : international medical journal of experimental and clinical research, 25, 2609–2622. https://doi.org/10.12659/MSM.913471

Chen, Y., & Wang, X. (2020). MiRDB: An online database for prediction of functional microRNA targets. Nucleic Acids Research, 48(D1), D127-D131. https://doi.org/10.1093/nar/gkz757

Chou K-C. Graphic Rule for Drug Metabolism Systems. Curr Drug Metab. 2010;11(4). doi:10.2174/138920010791514261

Cicenas J, Tamosaitis L, Kvederaviciute K, et al. KRAS, NRAS and BRAF mutations in colorectal cancer and melanoma. Med Oncol. 2017;34(2):1-11. doi:10.1007/s12032-016-0879-9

Cohen PA, Jhingran A, Oaknin A, Denny L. Cervical cancer. Lancet. 2019;393(10167):169-182. doi:10.1016/S0140-6736(18)32470-X

Cooke SL, Temple J, MacArthur S, et al. Intra-tumour genetic heterogeneity and poor chemoradiotherapy response in cervical cancer. Br J Cancer. 2011;104(2). doi:10.1038/sj.bjc.6605971

Crosbie EJ, Einstein MH, Franceschi S, Kitchener HC. Human papillo- mavirus and cervical cancer. Lancet. 2013;382:889-899. https://doi. org/10.1016/S0140-6736(13)60022-7

Csardi G, Nepusz T. The igraph software package for complex network research. InterJournal Complex Syst. 2006;Complex Sy(1695).

Daigo K, Takano A, Thang PM, et al. Characterization of KIF11 as a novel prognostic biomarker and therapeutic target for oral cancer. Int J Oncol. 2018;52(1). doi:10.3892/ijo.2017.4181

De Braekeleer E, Douet-Guilbert N, De Braekeleer M. RARA fusion genes in acute promyelocytic leukemia: A review. Expert Rev Hematol. 2014;7(3):347-357. doi:10.1586/17474086.2014.903794

Den Boon JA, Pyeon D, Wang SS, et al. Molecular transitions from papillomavirus infection to cervical precancer and cancer: Role of stromal estrogen receptor signaling. Proc Natl Acad Sci U S A. 2015;112(25). doi:10.1073/pnas.1509322112

Derrien, T., Johnson, R., Bussotti, G., Tanzer, A., Djebali, S., Tilgner, H., ... & Guigó, R. (2012). The GENCODE v7 catalog of human long noncoding RNAs: analysis of their gene structure, evolution, and expression. Genome research, 22(9), 1775-1789.

Deschavanne PJ, Giron A, Vilain J, Fagot G, Fertit B. Genomic signature: Characterization and classification of species assessed by chaos game representation of sequences. Mol BiolEvol. 1999;16(10). doi:10.1093/oxfordjournals.molbev.a026048

Du, H., & Che, G. (2017). Genetic alterations and epigenetic alterations of cancer-associated fibroblasts. Oncology letters, 13(1), 3-12.

Durinck S, Spellman PT, Birney E, Huber W. Mapping identifiers for the integration of genomic datasets with the R/ Bioconductor package biomaRt. Nat Protoc. 2009;4:1184-1191. https://doi.org/10. 1038/nprot.2009.97

Dutta C, Das J. Mathematical characterization of Chaos Game Representation. New algorithms for nucleotide sequence analysis. J Mol Biol. 1992;228(3). doi:10.1016/0022-2836(92)90857-G

Eissa S, Matboli M, Shehata HH, Essawy NOE. MicroRNA-10b and minichromosome maintenance complex component 5 gene as prognostic biomarkers in breast cancer. Tumor Biol. 2015;36(6). doi:10.1007/s13277-015-3090-2

Fan Y, Li H, Yu Z, et al. Long non-coding RNA FGD5-AS1 promotes non-small cell lung cancer cell proliferation through sponging hsa-miR-107 to up-regulate FGFRL1. Bioscience Reports. 2020 Jan;40(1):BSR20193309. DOI: 10.1042/bsr20193309. PMID: 31919528; PMCID: PMC6981095.

Ferenz NP, Gable A, Wadsworth P. Mitotic functions of kinesin-5. Semin Cell Dev Biol. 2010;21(3). doi:10.1016/j.semcdb.2010.01.019

Ferlay J, Ervik M, Lam F, et al. Global Cancer Observatory: Cancer Today. International Agency for Research on Cancer. (2018). https:// gco.iarc.fr/today.

Fu, X. D. (2014). Non-coding RNA: a new frontier in regulatory biology. National science review, 1(2), 190-204.

Gan, L., Yang, Y., Li, Q., Feng, Y., Liu, T., & Guo, W. (2018). Epigenetic regulation of cancer progression by EZH2: from biological insights to therapeutic potential. Biomarker Research, 6.

Gao W, Liu Y, Qin R, Liu D, Feng Q. Silence of fibronectin 1 increases cisplatin sensitivity of non-small cell lung cancer cell line. Biochem Biophys Res Commun. 2016;476(1). doi:10.1016/j.bbrc.2016.05.081

Gautier L, Cope L, Bolstad BM, Irizarry RA. Affy - Analysis of Affymetrix GeneChip data at the probe level. Bioinformatics. 2004;20(3). doi:10.1093/bioinformatics/btg405

Ghasemi O, Ma Y, Lindsey ML, Jin YF. Using systems biology approaches to understand cardiac inflammation and extracellular matrix remodeling in the setting of myocardial infarction. Wiley Interdiscip Rev Syst Biol Med. 2014;6(1). doi:10.1002/wsbm.1248

Ghosh A, Nandy A. Graphical representation and mathematical char- acterization of protein sequences and applications to viral proteins. Adv Protein Chem Struct Biol. 2011;83:1-42. https://doi.org/10.1016/ B978-0-12-381262-9.00001-X

Guvakova MA, Prabakaran I, Wu Z, et al. CDH2/N-cadherin and early diagnosis of invasion in patients with ductal carcinoma in situ. Breast Cancer Res Treat. 2020;183(2). doi:10.1007/s10549-020-05797-x

Győrffy, B. Discovery and ranking of the most robust prognostic biomarkers in serous ovarian cancer. GeroScience 45, 1889–1898 (2023). https://doi.org/10.1007/s11357-023-00742-4

Hajjari, M., & Salavaty, A. (2015). HOTAIR: an oncogenic long non-coding RNA in different cancers. Cancer biology & medicine, 12(1), 1–9. https://doi.org/10.7497/j.issn.2095-3941.2015.0006

Hao W, Yu M, Lin J, et al. The pan-cancer landscape of netrin family reveals potential oncogenic biomarkers. Sci Rep. 2020;10(1). doi:10.1038/s41598-020-62117-5

Hao, C., Lin, S., Liu, P., Liang, W., Li, Z., & Li, Y. (2023). Potential serum metabolites and long-chain noncoding RNA biomarkers for endometrial cancer tissue. The journal of obstetrics and gynaecology research, 49(2), 725–743. https://doi.org/10.1111/jog.15494

He PA, Wei J, Yao Y, Tie Z. A novel graphical representation of pro- teins and its application. Phys A Stat Mech Appl. 2012;391:93-99. https://doi.org/10.1016/j.physa.2011.08.015

He Z, Wang X, Yang Z, et al. Expression and prognosis of CDC45 in cervical cancer based on the GEO database. PeerJ. 2021;9. doi:10.7717/peerj.12114

Hejmadi, M. (2013). Introduction to Cancer Biology. (2 ed.) Ventus Publishing.

Hindumathi V, Kranthi T, Rao SB, Manimaran P. The prediction of candidate genes for cervix related cancer through gene ontology and graph theoretical approach. Mol Biosyst. 2014;10:1450-1460. https:// doi.org/10.1039/c4mb00004h

Hou W, Pan Q, He M. A new graphical representation of protein sequences and its applications. Phys A Stat Mech Appl. 2016;444:996- 1002. https://doi.org/10.1016/j.physa.2015.10.067

Hua M, Qin Y, Sheng M, et al. MiR-145 suppresses ovarian cancer progression via modulation of cell growth and invasion by targeting CCND2 and E2F3. Mol Med Rep. 2019;49:3575-3583. https://doi. org/10.3892/mmr.2019.10004

Huang H, Yang Y, Zhang W, Liu X, Yang G. TTK regulates proliferation and apoptosis of gastric cancer cells through the Akt-mTOR pathway. FEBS Open Bio. 2020;10(8). doi:10.1002/2211-5463.12909

Huang Y, Hu Y, Jin Z, Shen Z. LncRNA snaR upregulates GRB2-associated binding protein 2 and promotes proliferation of ovarian carcinoma cells. Biochem Biophys Res Commun. 2018;503: 2028-2032. https://doi.org/10.1016/j.bbrc.2018.07.152

Huang, R., Ding, P., & Yang, F. (2015). Clinicopathological significance and potential drug target of CDH1 in breast cancer: a meta-analysis and literature review. Drug Design, Development and Therapy, 9, 5277 - 5285.

Huang, X., Liu, X., Du, B., Liu, X., Xue, M., Yan, Q., ... & Wang, Q. (2021). LncRNA LINC01305 promotes cervical cancer progression through KHSRP and exosome-mediated transfer. Aging (Albany NY), 13(15), 19230.

Huang, Y., Lin, D., Cui, S., Huang, Y., Tang, Y., Xu, J., Bao, J., Li, Y., Wen, J., Zuo, H., Wang, W., Li, J., Ni, J., Ruan, Y., Li, L., Chen, Y., Xie, Y., Zhu, Z., Cai, X., . . . Huang, D. (2021). MiRTarBase update 2022: An informative resource for experimentally validated miRNA–target interactions. Nucleic Acids Research, 50(D1), D222. https://doi.org/10.1093/nar/gkab1079

Huarte, M. (2015). The emerging role of lncRNAs in cancer. Nature medicine, 21(11), 1253-1261.

Ijaz M, Wang F, Shahbaz M, Jiang W, Fathy AH, Nesa EU. The role of Grb2 in cancer and peptides as Grb2 antagonists. Protein Pept Lett. 2017;24: 1084-1095. https://doi.org/10.2174/0929866525666171123213148

Imai T, Oue N, Nishioka M, et al. Overexpression of KIF11 in gastric cancer with intestinal mucin phenotype. Pathobiology. 2016;84(1). doi:10.1159/000447303

Jeffrey HJ. Chaos game representation of gene structure. Nucleic Acids Res. 1990;18(8). doi:10.1093/nar/18.8.2163

Jeggari, A., Marks, D. S., & Larsson, E. (2012). MiRcode: A map of putative microRNA target sites in the long non-coding transcriptome. Bioinformatics, 28(15), 2062-2063. https://doi.org/10.1093/bioinformatics/bts344

Jeong H, Mason SP, Barabási AL, Oltvai ZN. Lethality and centrality in protein networks. Nature. 2001 May 3;411(6833):41-2. doi: 10.1038/35075138.

Jiang X, Feng L, Dai B, Li L, Lu W. Identification of key genes involved in nasopharyngeal carcinoma. Braz J Otorhinolaryngol. 2017;83(6). doi:10.1016/j.bjorl.2016.09.003

Jiang, X., Lai, Y., Wang, Q., Hu, Y., Yang, F., & Fan, H. (2020). Diagnostic and prognostic role of long noncoding RNA MALAT1 in breast cancer: a meta-analysis. Cancer Advances. DOI:10.53388/tmrc201800090

Jiang, Z. Q., & Zhou, W. X. (2011). Multifractal detrending moving-average cross-correlation analysis. Physical Review E, 84(1), 016106.

Jo, H., Shim, K., & Jeoung, D. (2022). Potential of the miR-200 family as a target for developing anti-cancer therapeutics. International Journal of Molecular Sciences, 23(11), 5881.

Ju LG, Zhu Y, Long QY, et al. SPOP suppresses prostate cancer through regulation of CYCLIN E1 stability. Cell Death Differ. 2019;26: 1156-1168. https://doi.org/10.1038/s41418-018-0198-0

Kaistha BP, Honstein T, Müller V, et al. Key role of dual specificity kinase TTK in proliferation and survival of pancreatic cancer cells. Br J Cancer. 2014;111(9). doi:10.1038/bjc.2014.460

Kallen, A. N., Zhou, X. B., Xu, J., Qiao, C., Ma, J., Yan, L., ... & Huang, Y. (2013). The imprinted H19 lncRNA antagonizes let-7 microRNAs. Molecular cell, 52(1), 101-112.

Kanbar, J. N., Ma, S., Kim, E. S., Kurd, N. S., Tsai, M. S., Tysl, T., Widjaja, C. E., Limary, A. E., Yee, B., He, Z., Hao, Y., Fu, X. D., Yeo, G. W., Huang, W. J., & Chang, J. T. (2022). The long noncoding RNA Malat1 regulates CD8+ T cell differentiation by mediating epigenetic repression. The Journal of experimental medicine, 219(6), e20211756. https://doi.org/10.1084/jem.20211756

Kantelhardt, J. W., Zschiegner, S. A., Koscielny-Bunde, E., Havlin, S., Bunde, A., & Stanley, H. E. (2002). Multifractal detrended fluctuation analysis of nonstationary time series. *Physica A: Statistical Mechanics and its Applications*, *316*(1-4), 87-114.

Kao SH, Wu HT, Wu KJ. Ubiquitination by HUWE1 in tumorigenesis and beyond. J Biomed Sci. 2018;25(1). doi:10.1186/s12929-018-0470-0

Kato T, Lee D, Wu L, et al. Kinesin family members KIF11 and KIF23 as potential therapeutic targets in malignant pleural mesothelioma. Int J Oncol. 2016;49(2). doi:10.3892/ijo.2016.3566

Kaushal, P., Singh, S. Network-based disease gene prioritization based on Protein–Protein Interaction Networks. Netw Model Anal Health Inform Bioinforma 9, 55 (2020). https://doi.org/10.1007/s13721-020-00260-9

Ke H, Kazi JU, Zhao H, Sun J. Germline mutations of KIT in gastrointestinal stromal tumor (GIST) and mastocytosis. Cell Biosci. 2016;6(1). doi:10.1186/s13578-016-0120-8

Kerrien S, Aranda B, Breuza L, et al. The IntAct molecular interaction database in 2012. Nucleic Acids Res. 2012;40(D1). doi:10.1093/nar/gkr1088

Khanna V, Pierce ST, Dao K-HT, et al. Durable Disease Control with MEK Inhibition in a Patient with NRAS-mutated Atypical Chronic Myeloid Leukemia. Cureus. 2015;7(12). doi:10.7759/cureus.414

Klein, G., Vellenga, E., Fraaije, M. W., Kamps, W. A., & De Bont, E. S. J. M. (2004). The possible role of matrix metalloproteinase (MMP)-2 and MMP-9 in cancer, eg acute leukemia. Critical reviews in oncology/hematology, 50(2), 87-100.

Kotlyar M, Pastrello C, Sheahan N, Jurisica I. Integrated interactions database: Tissue-specific view of the human and model organism interactomes. Nucleic Acids Res. 2016;44(D1). doi:10.1093/nar/gkv1115

Kulasingam V, Diamandis EP. Strategies for discovering novel cancer biomarkers through utilization of emerging technologies. Nat Clin Pract Oncol. 2008;5(10). doi:10.1038/ncponc1187

Kung, J. T., Colognori, D., & Lee, J. T. (2013). Long noncoding RNAs: past, present, and future. Genetics, 193(3), 651-669.

Kuo PL, Huang YL, Hsieh CCJ, Lee JC, Lin BW, Hung LY. STK31 Is a Cell-Cycle Regulated Protein That Contributes to the Tumorigenicity of Epithelial Cancer Cells. PLoS One. 2014;9(3). doi:10.1371/journal.pone.0093303

Lánczky A, Győrffy B. Web-based survival analysis tool tailored for medical research (KMplot): Development and implementation. J Med Internet Res. 2021;23(7). doi:10.2196/27633

Langfelder, P., & Horvath, S. (2008). WGCNA: An R package for weighted correlation network analysis. BMC Bioinformatics, 9. https://doi.org/10.1186/1471-2105-9-559

Lee, K. J., Singh, N., Bizuneh, M., Kim, N. H., Kim, H. S., Kim, Y., Lee, J. J., Kim, J. H., Kim, J., Jeong, S. Y., Cho, H. Y., & Park, S. T. (2023). miR-429 Suppresses Endometrial Cancer Progression and Drug Resistance via DDX53. Journal of personalized medicine, 13(9), 1302. https://doi.org/10.3390/jpm13091302

Lee, T. I., & Young, R. A. (2013). Transcriptional regulation and its misregulation in disease. Cell, 152(6), 1237-1251.

Levidou G, Ventouri K, Nonni A, et al. Replication protein A in nonearly ovarian adenocarcinomas: Correlation with MCM-2, MCM-5, Ki-67 index and prognostic significance. Int J Gynecol Pathol. 2012;31(4). doi:10.1097/PGP.0b013e31823ef92e

Li C, Xing L, Wang X. 2-D graphical representation of protein sequences and its application to coronavirus phylogeny. J Biochem Mol Biol. 2008;41:217-222. https://doi.org/10.5483/bmbrep.2008. 41.3.217

Li C, Yu X, Yang L, Zheng X, Wang Z. 3-D maps and coupling numbers for protein sequences. Phys A Stat Mech Appl. 2009;388:1967-1972. https://doi.org/10.1016/j.physa.2009.01.017

Li L, Ying J, Li H, et al. The human cadherin 11 is a pro-apoptotic tumor suppressor modulating cell stemness through Wnt/B-catenin signaling and silenced in common carcinomas. Oncogene. 2012;31(34). doi:10.1038/onc.2011.541

Li N, Piao J, Wang X, et al. Paip1 indicated poor prognosis in cervical cancer and promoted cervical carcinogenesis. Cancer Res Treat. 2019; 51:1653-1665.

Li X, Tian R, Gao H, et al. Identification of significant gene signatures and prognostic biomarkers for patients with cervical cancer by integrated bioinformatic methods. Technol Cancer Res Treat. 2018;17:153303381876745. doi:10.1177/1533033818767455

Li, L., Peng, Q., Gong, M., Ling, L., Xu, Y., & Liu, Q. (2021). Using lncRNA sequencing to reveal a putative lncRNA-mRNA correlation network and the potential role of PCBP1-AS1 in the pathogenesis of cervical cancer. Frontiers in Oncology, 11, 634732.

Li, Z., Ivanov, A. A., Su, R., Gonzalez-Pecchi, V., Qi, Q., Liu, S., ... & Fu, H. (2017). The OncoPPi network of cancer-focused protein–protein interactions to inform biological insights and therapeutic strategies. Nature communications, 8(1), 14356.

Liang B, Li C, Zhao J. Identification of key pathways and genes in colorectal cancer using bioinformatics analysis. Med Oncol. 2016;33(10). doi:10.1007/s12032-016-0829-6

Liao B, Liao B, Sun X, Zeng Q. A novel method for similarity analysis and protein sub-cellular localization prediction. Bioinformatics. 2010; 26:2678-2683. https://doi.org/10.1093/bioinformatics/btq521

Licata L, Briganti L, Peluso D, et al. MINT, the molecular interaction database: 2012 Update. Nucleic Acids Res. 2012;40(D1). doi:10.1093/nar/gkr930

Lim G, Huh WK. Rad52 phosphorylation by Ipl1 and Mps1 contributes to Mps1 kinetochore localization and spindle assembly checkpoint regulation. Proc Natl Acad Sci U S A. 2017;114(44). doi:10.1073/pnas.1705261114

Lin, C. P., & He, L. (2017). Noncoding RNAs in cancer development. Annual review of cancer biology, 1, 163-184.

Ling, Z., Li, P., Ge, M., Zhao, X., Hu, F., Fang, X., Dong, Z., & Mao, W. (2011). Hypermethylation-modulated down-regulation of CDH1 expression contributes to the progression of esophageal cancer. International journal of molecular medicine, 27 5, 625-35 .

Liu YQ, Zou HY, Xie JJ, Fang WK. Paradoxical roles of desmosomal components in head and neck cancer. Biomolecules. 2021;11(6). doi:10.3390/biom11060914

Liu, T., Wang, H., Yu, H., Bi, M., Yan, Z., Hong, S., & Li, S. (2020). The Long Non-coding RNA HOTTIP Is Highly Expressed in Colorectal Cancer and Enhances Cell Proliferation and Invasion. Molecular therapy. Nucleic acids, 19, 612–618. https://doi.org/10.1016/j.omtn.2019.12.008

Löchel HF, Heider D. Chaos game representation and its applications in bioinformatics. Comput Struct Biotechnol J. 2021;19:6263-6271

Lou X, Han X, Jin C, et al. SOX2 targets fibronectin 1 to promote cell migration and invasion in ovarian cancer: New molecular leads for therapeutic intervention. Omi A J Integr Biol. 2013;17(10):510-518. doi:10.1089/omi.2013.0058

Love, M. I., Huber, W., & Anders, S. (2014). Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. Genome Biology, 15(12). https://doi.org/10.1186/s13059-014-0550-8

Lu JL, Hu XH, Liu X, Shi F. Predicting thermophilic nucleotide sequences based on chaos game representation features and support vector machine. In: 5th International Conference on Bioinformatics and Biomedical Engineering, ICBBE 2011; 2011. doi:10.1109/icbbe.2011.5780070

Lu S, Qian J, Guo M, Gu C, Yang Y. Insights into a Crucial Role of TRIP13 in Human Cancer. Comput Struct Biotechnol J. 2019;17:854-861. doi:10.1016/j.csbj.2019.06.005

Luo, Y., Riedlinger, G., & Szolovits, P. (2014). Text mining in cancer gene and pathway prioritization. Cancer informatics, 13(Suppl 1), 69–79. https://doi.org/10.4137/CIN.S13874

Ma, C., Nong, K., Zhu, H., Wang, W., Huang, X., Yuan, Z., & Ai, K. (2014). H19 promotes pancreatic cancer metastasis by derepressing let-7's suppression on its target HMGA2-mediated EMT. Tumor Biology, 35, 9163-9169.

Mandelbrot B. B. The fractal geometry of nature /Revised and enlarged edition/. New York. Published online 1983.

Manimaran, P., Panigrahi, P. K., & Parikh, J. C. (2005). Wavelet analysis and scaling properties of time series. Physical Review E, 72(4), 046120.

Mao Y, Schwarzbauer JE. Fibronectin fibrillogenesis, a cell-mediated matrix assembly process. Matrix Biol. 2005;24(6). doi:10.1016/j.matbio.2005.06.008

Masai H, You Z, Arai KI. Control of DNA replication: Regulation and activation of eukaryotic replicative helicase, MCM. IUBMB Life. 2005;57(4-5). doi:10.1080/15216540500092419

Mattick, J. S., & Rinn, J. L. (2015). Discovery and annotation of long noncoding RNAs. Nature structural & molecular biology, 22(1), 5-7.

McDowall MD, Scott MS, Barton GJ. PIPs: Human protein-protein interaction prediction database. Nucleic Acids Res. 2009;37(SUPPL. 1). doi:10.1093/nar/gkn870

McGeary SE, Lin KS, Shi CY, Pham T, Bisaria N, Kelley GM, Bartel DP. The biochemical basis of microRNA targeting efficacy. Science Dec 5, (2019).

Miki T, Smith CL, Long JE, Eva A, Fleming TP. Oncogene ect2 is related to regulators of small GTP-binding proteins. Nature. 1993;362(6419). doi:10.1038/362462a0

Milenkovic, T., Memisevic, V., Ganesan, A. K., & Przulj, N. (2010). Systems-level cancer gene identification from protein interaction network topology applied to melanogenesis-related functional genomics data. Journal of the Royal Society, Interface, 7(44), 423–437. https://doi.org/10.1098/rsif.2009.0192

Miniowitz-Shemtov S, Eytan E, Kaisari S, Sitry-Shevah D, Hershko A. Mode of interaction of TRIP13 AAA-ATPase with the Mad2-binding protein p31comet and with mitotic checkpoint complexes. Proc Natl Acad Sci U S A. 2015;112(37):11536-11540. doi:10.1073/pnas.1515358112

Nannini M, Pantaleo MA, Maleddu A, Astolfi A, Formica S, Biasco G. Gene expression profiling in colorectal cancer using microarray technologies: Results and perspectives. Cancer Treat Rev. 2009;35(3). doi:10.1016/j.ctrv.2008.10.006

Newman MEJ. Contemporary Physics Power laws, Pareto distributions and Zipf's law. Contemp Phys. 2005;46(5).

Nie, L., Wu, H. J., Hsu, J. M., Chang, S. S., Labaff, A. M., Li, C. W., Wang, Y., Hsu, J. L., & Hung, M. C. (2012). Long non-coding RNAs: versatile master regulators of gene expression and crucial players in cancer. American journal of translational research, 4(2), 127–150.

Oany AR, Mia M, Pervin T, Alyami SA, Moni MA. Integrative systems biology approaches to identify potential biomarkers and pathways of cervical cancer. J Pers Med. 2021;11(5). doi:10.3390/jpm11050363

Ohno, S. (1972). So much" junk" DNA in our genome. In" Evolution of Genetic Systems". In Brookhaven Symposium in Biology (Vol. 23, pp. 366-370).

Ooi A, Oyama T, Nakamura R, et al. Gene amplification of CCNE1, CCND1, and CDK6 in gastric cancers detected by multiplex ligation- dependent probe amplification and fluorescence in situ hybridization. Hum Pathol. 2017;61:58-67. https://doi.org/10.1016/j.humpath. 2016.10.025

Otálora-Otálora BA, Henríquez B, Lopez-Kleine L, Rojas A. RUNX family: oncogenes or tumor suppressors (review). Oncol Rep. 2019;42: 3-19. https://doi.org/10.3892/or.2019.7149

Oughtred R, Rust J, Chang C, et al. The BioGRID database: A comprehensive biomedical resource of curated protein, genetic, and chemical interactions. Protein Sci. 2021;30(1). doi:10.1002/pro.3978

Pagès H, Aboyoun P, Gentleman R, DebRoy S. Biostrings: Efficient manipulation of biological strings. R package version 2.46.0, R Packag. Version 2.46.0. 2017.

Pal M, Kiran VS, Rao PM, Manimaran P. Multifractal detrended cross-correlation analysis of genome sequences using chaos-game representation. Phys A Stat Mech its Appl. 2016;456. doi:10.1016/j.physa.2016.03.074

Pal M, Satish B, Srinivas K, Rao PM, Manimaran P. Multifractal detrended cross-correlation analysis of coding and non-coding DNA sequences through chaos-game representation. Phys A Stat Mech its Appl. 2015;436. doi:10.1016/j.physa.2015.05.018

Pal, M., & Manimaran, P. (2019). Multifractal detrended partial cross-correlation analysis on Asian markets. Physica A: Statistical Mechanics and its Applications, 531, 121778.

Pankov R, Yamada KM. Fibronectin at a glance. J Cell Sci. 2002;115(20). doi:10.1242/jcs.00059

Park SY, Lee CJ, Choi JH, et al. The JAK2/STAT3/CCND2 Axis pro- motes colorectal cancer stem cell persistence and radioresistance. J Exp Clin Cancer Res. 2019;38:399. https://doi.org/10.1186/s13046- 019-1405-7

Patil A, Nakai K, Nakamura H. HitPredict: A database of quality assessed protein-protein interactions in nine species. Nucleic Acids Res. 2011;39(SUPPL. 1). doi:10.1093/nar/gkq897

Perez-Iratxeta C, Wjst M, Bork P, Andrade MA. G2D: a tool for min- ing genes associated with disease. BMC Genet. 2005;6:45. https:// doi.org/10.1186/1471-2156-6-45

Piao J, Chen L, Jin T, Xu M, Quan C, Lin Z. Paip1 affects breast can- cer cell growth and represents a novel prognostic biomarker. Hum Pathol. 2018;73:33-40. https://doi.org/10.1016/j.humpath.2017. 10.037

Piñero J, Ramírez-Anguita JM, Saüch-Pitarch J, et al. The DisGeNET knowledge platform for disease genomics: 2019 update. Nucleic Acids Res. 2020;48(D1):D845-D855. doi:10.1093/nar/gkz1021

Piñero J, Ramírez-Anguita JM, Saüch-Pitarch J, et al. The DisGeNET knowledge platform for disease genomics: 2019 update. Nucleic Acids Res. 2020;48:D845-D855. https://doi.org/10.1093/nar/gkz1021

Podder A, Pandit M, Narayanan L. Drug target prioritization for Alzheimer's disease using protein interaction network analysis. Omi A J Integr Biol. 2018;22(10). doi:10.1089/omi.2018.0131

Podobnik, B., & Stanley, H. E. (2008). Detrended cross-correlation analysis: a new method for analyzing two nonstationary time series. Physical review letters, 100(8), 084102.

Podobnik, B., Horvatic, D., Petersen, A. M., & Stanley, H. E. (2009). Cross-correlations between volume change and price change. Proceedings of the National Academy of Sciences, 106(52), 22079-22084.

Pollok S, Bauerschmidt C, Sänger J, Nasheuer HP, Grosse F. Human Cdc45 is a proliferation-associated antigen. FEBS J. 2007;274(14). doi:10.1111/j.1742-4658.2007.05900.x

Qiao XR, Zhang X, Mu L, Tian J, Du Y. GRB2-associated binding pro- tein 2 regulates multiple pathways associated with the development of prostate cancer. Oncol Lett. 2020;20:99. https://doi.org/10.3892/ ol.2020.11960

Qin S, Gao Y, Yang Y, et al. Identifying molecular markers of cervical cancer based on competing endogenous RNA network analysis. Gynecol Obstet Invest. 2019;84:350-359. https://doi.org/10.1159/ 000493476

Qing S, Tulake W, Ru M, et al. Proteomic identification of potential biomarkers for cervical squamous cell carcinoma and human papillomavirus infection. Published online 2017. doi:10.1177/1010428317697547

Qiu HZ, Huang J, Xiang CC, et al. Screening and Discovery of New Potential Biomarkers and Small Molecule Drugs for Cervical Cancer: A Bioinformatics Analysis. Technol Cancer Res Treat. 2020;19. doi:10.1177/1533033820980112

R Core Team. R: A Language and Environment for Statistical Comput- ing (2020). https://www.r-project.org/.

Radha, G., & Raghavan, S.C. (2017). BCL2: A promising cancer therapeutic target. Biochimica et biophysica acta. Reviews on cancer, 1868 1, 309-314 .

Rafique M, Iqbal J, Lone KJ, Mir AA, Kearfott KJ, Iqbal A, Qureshi SA, Abbasi SA, Nikolopoulos D, khan TM. Multifractal detrended cross-correlation analysis of radioactivity borne radon, thoron and meteorological time series. Phys A Stat Mech its Appl. 2022; 607. https://doi.org/10.1016/j.physa.2022.128214

Raman K. Construction and analysis of protein-protein interaction networks. Autom Exp. 2010;2(1). doi:10.1186/1759-4499-2-2

Randic M, Butina D, Zupan J. Novel 2-D graphical representation of proteins. Chem Phys Lett. 2006;419:528-532. https://doi.org/10. 1016/j.cplett.2005.11.091

Randic M, Novicˇ M, Vracˇko M. On novel representation of proteins based on amino acid adjacency matrix. SAR QSAR Environ Res. 2008; 19:339-349. https://doi.org/10.1080/10629360802085082

Rao X, Wang J, Song HM, Deng B, Li JG. KRT15 overexpression pre- dicts poor prognosis in colorectal cancer. Neoplasma. 2020;67:410- 414. https://doi.org/10.4149/neo_2019_190531N475

Rao, X., & Lu, Y. (2022). C1QTNF6 Targeted by MiR-184 Regulates the Proliferation, Migration, and Invasion of Lung Adenocarcinoma Cells. Molecular biotechnology, 64(11), 1279–1287. https://doi.org/10.1007/s12033-022-00495-z

Rapley J, Nicolàs M, Groen A, et al. The NIMA-family kinase Nek6 phosphorylates the kinesin Eg5 at a novel site necessary for mitotic spindle formation. J Cell Sci. 2008;121(23). doi:10.1242/jcs.035360

Rappaport N, Twik M, Plaschkes I, et al. MalaCards: An amalgamated human disease compendium with diverse clinical and genetic annotation and structured search. Nucleic Acids Res. 2017;45(D1). doi:10.1093/nar/gkw1012

Rashid, F., Shah, A., & Shan, G. (2016). Long non-coding RNAs in the cytoplasm. Genomics, proteomics & bioinformatics, 14(2), 73-80.

Rasool, M., Malik, A., Zahid, S., Ashraf, M. A. B., Qazi, M. H., Asif, M., ... & Jamal, M. S. (2016). Non-coding RNAs in cancer diagnosis and therapy. Non-coding RNA research, 1(1), 69-76.

Reichheld A, Mukherjee PK, Rahman SMF, David K V., Pricilla RA. Prevalence of cervical cancer screening and awareness among women in an urban community in South India—a cross sectional study. Ann Glob Heal. 2020;86(1). doi:10.5334/aogh.2735

Ren L, Yi J, Li W, et al. Apolipoproteins and cancer. Cancer Med. 2019;8:7032-7043. https://doi.org/10.1002/cam4.2587

Repana D, Nulsen J, Dressler L, et al. The network of cancer genes (NCG): a comprehensive catalogue of known and candidate cancer genes from cancer sequencing screens. Genome Biol. 2019;20:1-12. https://doi.org/10.1186/s13059-018-1612-0

Rickman DS, Schulte JH, Eilers M. The expanding world of N-MYC- driven tumors. Cancer Discov. 2018;8:150-164. https://doi.org/10. 1158/2159-8290.CD-17-0273

Ritchie ME, Phipson B, Wu D, et al. Limma powers differential expression analyses for RNA-sequencing and microarray studies. Nucleic Acids Res. 2015;43(7). doi:10.1093/nar/gkv007

Rual JF, Venkatesan K, Hao T, et al. Towards a proteome-scale map of the human protein-protein interaction network. Nature. 2005;437(7062). doi:10.1038/nature04209

Sahoo GR, Dey R, Das N, Ghosh N, Pradhan A. Two dimensional multifractal detrended fluctuation analysis of low coherence images for diagnosis of cervical pre-cancer. Biomed Phys Eng Express. 2020;7(2). doi:10.1088/2057-1976/ab6e17

Saleembhasha, A., & Mishra, S. (2019). Long non-coding RNAs as pan-cancer master gene regulators of associated protein-coding genes: a systems biology approach. PeerJ, 7, e6388. https://doi.org/10.7717/peerj.6388

Salmena L, Poliseno L, Tay Y et al: A ceRNA hypothesis: The Rosetta Stone of a hidden RNA language? Cell, 2011; 146(3): 353–58

Sanchez-Solana B, Wang D, Qian X, et al. The tumor suppressor activity of DLC1 requires the interaction of its START domain with Phosphatidylserine, PLCD1, and Caveolin-1. Mol Cancer. 2021;20(1). doi:10.1186/s12943-021-01439-y

Satelli, A., Li, S. Vimentin in cancer and its potential as a molecular target for cancer therapy. Cell. Mol. Life Sci. 68, 3033–3046 (2011). https://doi.org/10.1007/s00018-011-0735-1

Schneider MA, Christopoulos P, Muley T, et al. AURKA, DLGAP5, TPX2, KIF11 and CKAP5: Five specific mitosis-associated genes correlate with poor prognosis for non-small cell lung cancer patients. Int J Oncol. 2017;50(2). doi:10.3892/ijo.2017.3834

Scotto L, Narayan G, Nandula S V., et al. Identification of copy number gain and overexpressed genes on chromosome arm 20q by an integrative genomic approach in cervical cancer: Potential role in progression. Genes Chromosom Cancer. 2008;47(9). doi:10.1002/gcc.20577

Sekino Y, Han X, Kobayashi G, et al. BUB1B Overexpression Is an Independent Prognostic Marker and Associated with CD44, p53, and PD-L1 in Renal Cell Carcinoma. Oncol. 2021;99(4). doi:10.1159/000512446

Shannon, P., Markiel, A., Ozier, O., Baliga, N. S., Wang, J. T., Ramage, D., Amin, N., Schwikowski, B., & Ideker, T. (2003). Cytoscape: A software Environment for integrated models of biomolecular interaction networks. Genome Research, 13(11). https://doi.org/10.1101/gr.1239303

Shen J, Li M, Min L. HSPB8 promotes cancer cell growth by activating the ERK-CREB pathway and is indicative of a poor prognosis in gas- tric cancer patients. Oncol Rep. 2018;39:2978-2986. https://doi.org/ 10.3892/or.2018.6376

Shen, L., Chen, L., Wang, Y., Jiang, X., Xia, H., & Zhuang, Z. (2015). Long noncoding RNA MALAT1 promotes brain metastasis by inducing epithelial-mesenchymal transition in lung cancer. Journal of neuro-oncology, 121(1), 101–108. https://doi.org/10.1007/s11060-014-1613-0

Shen, L., Cui, J., Liang, S., Pang, Y., & Liu, P. (2013). Update of research on the role of EZH2 in cancer progression. OncoTargets and therapy, 6, 321 - 324.

Shen, Z., Zhou, C., Li, J., Deng, H., Li, Q., & Wang, J. (2016). The association, clinicopathological significance, and diagnostic value of CDH1 promoter methylation in

head and neck squamous cell carcinoma: a meta-analysis of 23 studies. OncoTargets and therapy, 9, 6763 - 6773.

Silva RD, Mirkovic M, Guilgur LG, Rathore OS, Martinho RG, Oliveira RA. Absence of the Spindle Assembly Checkpoint Restores Mitotic Fidelity upon Loss of Sister Chromatid Cohesion. Curr Biol. 2018;28(17). doi:10.1016/j.cub.2018.06.062

Song G, Liu K, Yang X, et al. SATB1 plays an oncogenic role in esophageal cancer by upregulation of FN1 and PDGFRB. Oncotarget. 2017;8(11). doi:10.18632/oncotarget.14849

Song M, Wang Y, Zhang Z, Wang S. PSMC2 is up-regulated in osteosarcoma and regulates osteosarcoma cell proliferation, apoptosis and migration. Oncotarget. 2017;8(1). doi:10.18632/oncotarget.13511

Steinwald P, Ledet E, Sartor O. Eradication of BRAF K601E Mutation in Metastatic Castrate-resistant Prostate Cancer Treated With Cabazitaxel and Carboplatin: A Case Report. Clin Genitourin Cancer. 2020;18(3):e312-e314. doi:10.1016/j.clgc.2019.12.015

Stelzl U, Worm U, Lalowski M, et al. A human protein-protein interaction network: A resource for annotating the proteome. Cell. 2005;122(6). doi:10.1016/j.cell.2005.08.029

Strouhalova, K., Přechová, M., Gandalovičová, A., Brábek, J., Gregor, M., & Rosel, D. (2020). Vimentin intermediate filaments as potential target for cancer treatment. Cancers, 12(1), 184.

Sun, L., Luo, H., Liao, Q., Bu, D., Zhao, G., Liu, C., ... & Zhao, Y. (2013). Systematic study of human long intergenic non-coding RNAs and their impact on cancer. Science China Life Sciences, 56, 324-334.

Sun, Q., Hao, Q., & Prasanth, K. V. (2018). Nuclear long noncoding RNAs: key regulators of gene expression. Trends in Genetics, 34(2), 142-157.

Sun, W., Yang, Y., Xu, C., & Guo, J. (2017). Regulatory mechanisms of long noncoding RNAs on gene expression in cancers. Cancer genetics, 216-217, 105–110. https://doi.org/10.1016/j.cancergen.2017.06.003

Sung H, Ferlay J, Siegel RL, et al. Global Cancer Statistics 2020: GLOBOCAN Estimates of Incidence and Mortality Worldwide for 36 Cancers in 185 Countries. CA Cancer J Clin. 2021;71(3). doi:10.3322/caac.21660

Szklarczyk D, Gable AL, Nastou KC, et al. Erratum: Correction to “The STRING database in 2021: customizable protein-protein networks, and functional characterization of user-

uploaded gene/measurement sets" (Nucleic acids research (2021) 49 D1 (D605-D612)). Nucleic Acids Res. 2021;49(18). doi:10.1093/nar/gkab835

Tanchotsrinon W, Lursinsap C, Poovorawan Y. A high performance prediction of HPV genotypes by Chaos game representation and singular value decomposition. BMC Bioinformatics. 2015;16(1). doi:10.1186/s12859-015-0493-4

Terai, G., Iwakiri, J., Kameda, T. et al. Comprehensive prediction of lncRNA–RNA interactions in human transcriptome. BMC Genomics 17 (Suppl 1), 12 (2016). https://doi.org/10.1186/s12864-015-2307-5

Thippana, M., Dwivedi, A., Palanisamy, M., & Vindal, V. (2023). Identification of key molecular players and associated pathways in cervical squamous cell carcinoma progression through network analysis. September 2022, 1–15. https://doi.org/10.1002/prot.26502

Thummadi NB, Charutha S, Pal M, Manimaran P. Multifractal and cross-correlation analysis on mitochondrial genome sequences using chaos game representation. Mitochondrion. 2021;60. doi:10.1016/j.mito.2021.08.006

Tian, Q., Yan, X., Yang, L., Liu, Z., Yuan, Z., & Zhang, Y. (2021). lncRNA CYTOR promotes cell proliferation and tumor growth via miR-125b/SEMA4C axis in hepatocellular carcinoma. Oncology letters, 22(5), 796. https://doi.org/10.3892/ol.2021.13057

Tranchevent, L. C., Capdevila, F. B., Nitsch, D., De Moor, B., De Causmaecker, P., & Moreau, Y. (2011). A guide to web tools to prioritize candidate genes. Briefings in bioinformatics, 12(1), 22–32. https://doi.org/10.1093/bib/bbq007

Travasso CM, Anand M, Samarth M, Deshpande A, Kumar-Sinha C. Human papillomavirus genotyping by multiplex pyrosequencing in cervical cancer patients from India. J Biosci. 2008;33(1). doi:10.1007/s12038-008-0023-x

Turner FS, Clutterbuck DR, Semple CA. Article R75, BioMed Central, 2003. Accessed February 21, 2021. http://genomebiology.com/ 2003/4/11/R75

Tyanova S, Cox J. Perseus: a bioinformatics platform for integrative analysis of proteomics data in cancer research. In: von Stechow L, ed. Cancer Systems Biology. Methods in Molecular Biology. Springer New York; 2018:133-148. https://doi.org/10.1007/978-1-4939- 7493-1_7

Uhlén M, Fagerberg L, Hallström BM, et al. Tissue-based map of the human proteome. Science (80-). 2015;347(6220). doi:10.1126/science.1260419

Vader G. Pch2TRIP13: controlling cell division through regulation of HORMA domains. Chromosoma. 2015;124(3):333-339. doi:10.1007/s00412-015-0516-y

Vendramini E, Bomben R, Pozzo F, et al. KRAS, NRAS, and BRAF mutations are highly enriched in trisomy 12 chronic lymphocytic leukemia and are associated with shorter treatment-free survival. Leukemia. 2019;33(8):2111-2115. doi:10.1038/s41375-019-0444-6

Venere M, Horbinski C, Crish JF, et al. The mitotic kinesin KIF11 is a driver of invasion, proliferation, and self-renewal in glioblastoma. Sci Transl Med. 2015;7(304). doi:10.1126/scitranslmed.aac6762

Vidal, M., Cusick, M. E., & Barabási, A. L. (2011). Interactome Networks and Human Disease. Cell, 144(6), 986–998. https://doi.org/10.1016/J.CELL.2011.02.016

Viechtbauer, W. (2010). Conducting meta-analyses in R with the metafor package. Journal of statistical software, 36, 1-48.

Wang K, Sturt-Gillespie B, Hittle JC, et al. Thyroid hormone receptor interacting protein 13 (TRIP13) AAA-ATPase is a novel mitotic checkpoint-silencing protein. J Biol Chem. 2014;289(34). doi:10.1074/jbc.M114.585315

Wang S, Wu Z, Li T, et al. Mutational spectrum and prognosis in NRAS-mutated acute myeloid leukemia. Sci Rep. 2020;10(1):12152. doi:10.1038/s41598-020-69194-6

Wang, J. Y., & Chen, L. J. (2019). The role of miRNAs in the invasion and metastasis of cervical cancer. Bioscience reports, 39(3), BSR20181377.

Warde N. Prostate cancer: Cadherin 2: An important new player in castration resistance. Nat Rev Urol. 2011;8(1). doi:10.1038/nrurol.2010.222

Wąż P, Bielinska-Wąż D. Moments of inertia of spectra and distribu- tion moments as molecular descriptors. MATCH Commun Math Com- put Chem. 2013;70:851-865.

Wei J, Wang Y, Shi K, Wang Y. Identification of Core prognosis- related candidate genes in cervical cancer via integrated bioinformatical analysis. Biomed Res Int. 2020;2020:1-9. https://doi. org/10.1155/2020/8959210

Wen J, Zhang YY. A 2D graphical representation of protein sequence and its numerical characterization. Chem Phys Lett. 2009;476:281- 286. https://doi.org/10.1016/j.cplett.2009.06.017

Wong YF, Selvanayagam ZE, Wei N, et al. Expression Genomics of Cervical Cancer: Molecular Classification and Prediction of Radiotherapy Response by DNA Microarray. Clin Cancer Res. 2003;9(15).

Wu ZC, Xiao X, Chou KC. 2D-MH: A web-server for generating graphic representation of protein sequences based on the physicochemical properties of their constituent amino acids. J Theor Biol. 2010;267(1).

Wu, G., Liu, J., Wu, Z., Wu, X., & Yao, X. (2017). MicroRNA-184 inhibits cell proliferation and metastasis in human colorectal cancer by directly targeting IGF-1R. Oncology letters, 14(3), 3215–3222. https://doi.org/10.3892/ol.2017.6499

Wu, T., Hu, E., Xu, S., Chen, M., Guo, P., Dai, Z., Feng, T., Zhou, L., Tang, W., Zhan, L., Fu, X., Liu, S., Bo, X., & Yu, G. (2021). ClusterProfiler 4.0: A universal enrichment tool for interpreting omics data. The Innovation, 2(3), 100141. https://doi.org/10.1016/j.xinn.2021.100141

Xenarios I, Rice DW, Salwinski L, Baron MK, Marcotte EM, Eisenberg D. DIP: The Database of Interacting Proteins. Nucleic Acids Res. 2000;28(1). doi:10.1093/nar/28.1.289

Xiao Q, Zhou J, Shi L. A novel 3D graphical representation of RNA secondary structures based on chaos game representation. In: Proceedings - 2010 6th International Conference on Natural Computation, ICNC 2010. Vol 6. ; 2010. doi:10.1109/ICNC.2010.5582459

Xu, J., Wang, X., Zhu, C., & Wang, K. (2022). A review of current evidence about lncRNA MEG3: A tumor suppressor in multiple cancers. Frontiers in cell and developmental biology, 10, 997633. https://doi.org/10.3389/fcell.2022.997633

Xu, W. W., Jin, J., Wu, X. Y., Ren, Q. L., & Farzaneh, M. (2022). MALAT1-related signaling pathways in colorectal cancer. Cancer cell international, 22(1), 126. https://doi.org/10.1186/s12935-022-02540-y

Xue S, Jiang SQ, Li QW, et al. Decreased expression of BRAF- activated long non-coding RNA is associated with the proliferation of clear cell renal cell carcinoma. BMC Urol. 2018;18:79. https://doi.org/ 10.1186/s12894-018-0395-7

Yan, L., Zhou, J., Gao, Y., Ghazal, S., Lu, L., Bellone, S., ... & Huang, Y. (2015). Regulation of tumor cell migration and invasion by the H19/let-7 axis is antagonized by metformin-induced DNA methylation. Oncogene, 34(23), 3076-3084.

Yang L, Yang Y, Meng M, et al. Identification of prognosis-related genes in the cervical cancer immune microenvironment. Gene. 2021; 766:145119. https://doi.org/10.1016/j.gene.2020.145119

Yang Q, Huo S, Sui Y, et al. Mutation status and immunohistochemical correlation of KRAS, NRAS, and BRAF in 260 Chinese colorectal and gastric cancers. Front Oncol. 2018;8:487. https://doi.org/10.3389/ fonc.2018.00487

Yang, Y., Han, L., Yuan, Y., Li, J., Hei, N., & Liang, H. (2014). Gene co-expression network analysis reveals common system-level properties of prognostic genes across cancer types. Nature communications, 5, 3231. https://doi.org/10.1038/ncomms4231

Ye B, Duan B, Deng W, et al. EGF stimulates Rab35 activation and gastric cancer cell migration by regulating DENND1A-Grb2 complex formation. Front Pharmacol. 2018;9:1343. https://doi.org/10.3389/ fphar.2018.01343

Yi L, Lei Y, Yuan F, Tian C, Chai J, Gu M. NTN4 as a prognostic marker and a hallmark for immune infiltration in breast cancer. Sci Rep. 2022;12(1). doi:10.1038/s41598-022-14575-2

Yokoe T, Tanaka F, Mimori K, et al. Efficient identification of a novel cancer/testis antigen for immunotherapy using three-step microarray analysis. Cancer Res. 2008;68(4). doi:10.1158/0008-5472.CAN-07-0964

Yost S, De Wolf B, Hanks S, et al. Biallelic TRIP13 mutations predispose to Wilms tumor and chromosome missegregation. Nat Genet. 2017;49(7). doi:10.1038/ng.3883

You W, Henneberg M. Cancer incidence increasing globally: The role of relaxed natural selection. Evol Appl. 2018;11(2). doi:10.1111/eva.12523

Yu G, Wang LG, Han Y, He QY. ClusterProfiler: An R package for comparing biological themes among gene clusters. Omi A J Integr Biol. 2012;16(5). doi:10.1089/omi.2011.0118

Yu H, Kim PM, Sprecher E, Trifonov V, Gerstein M. The importance of bottlenecks in protein networks: Correlation with gene essentiality and expression dynamics. PLoS Comput Biol. 2007;3(4). doi:10.1371/journal.pcbi.0030059

Yu SY, Wang YP, Chang JYF, Shen WR, Chen HM, Chiang CP. Increased expression of MCM5 is significantly associated with aggressive progression and poor prognosis of oral squamous cell carcinoma. J Oral Pathol Med. 2014;43(5). doi:10.1111/jop.12134

Yu ZG, Anh V, Lau KS. Chaos game representation of protein sequences based on the detailed HP model and their multifractal and correlation analyses. J Theor Biol. 2004;226(3). doi:10.1016/j.jtbi.2003.09.009

Yuan Y, Shi X, Li B, et al. Integrated analysis of key microRNAs /TFs /mRNAs/ in HPV-positive cervical cancer based on microRNA sequencing and bioinformatics analysis. Pathol Res Pract. 2020;216: 152952. https://doi.org/10.1016/j.prp.2020.152952

Zanzoni, A., Soler-López, M., & Aloy, P. (2009). A network medicine approach to human disease. FEBS Letters, 583(11), 1759–1765. https://doi.org/10.1016/J.FEBSLET.2009.03.001

Zeng R, Liu Y, Jiang ZJ, et al. EPB41L3 is a potential tumor suppressor gene and prognostic indicator in esophageal squamous cell carcinoma. Int J Oncol. 2018;52(5). doi:10.3892/ijo.2018.4316

Zhai Y, Kuick R, Nan B, et al. Gene expression analysis of preinvasive and invasive cervical squamous cell carcinomas identifies HOXC10 as a key mediator of invasion. Cancer Res. 2007;67(21). doi:10.1158/0008-5472.CAN-07-2056

Zhang P, Zhang J, Sheng H, Russo JJ, Osborne B, Buetow K. Gene functional similarity search tool (GFSST). BMC Bioinform. 2006;7:135. https://doi.org/10.1186/1471-2105-7-135

Zhang Y, Yan WT, Yang ZY, et al. The role of WT1 in breast cancer: clinical implications, biological effects and molecular mechanism. Int J Biol Sci. 2020;16:1474-1480. https://doi.org/10.7150/ijbs. 39958

Zhang Z, Luo Y, Hu S, Li X, Wang L, Zhao B. A novel method to predict essential proteins based on tensor and HITS algorithm. Hum Genomics. 2020;14:14. https://doi.org/10.1186/s40246-020-00263-7

Zhang, Y., Xu, L., Li, A., & Han, X. (2019). The roles of ZEB1 in tumorigenic progression and epigenetic modifications. Biomedicine & Pharmacotherapy, 110, 400-408.

Zhao, H., Yin, X., Xu, H., Liu, K., Liu, W., Wang, L., Zhang, C., Bo, L., Lan, X., Lin, S., Feng, K., Ning, S., Zhang, Y., & Wang, L. (2023). LncTarD 2.0: An updated comprehensive database for experimentally-supported functional lncRNA–target regulations in human diseases. Nucleic Acids Research, 51(D1), D199-D207. https://doi.org/10.1093/nar/gkac984

Zheng, Q., Zhang, Q., Yu, X., He, Y., & Guo, W. (2021). FENDRR: A pivotal, cancer-related, long non-coding RNA. Biomedicine & pharmacotherapy = Biomedecine & pharmacotherapie, 137, 111390. https://doi.org/10.1016/j.biopha.2021.111390

Zhou GP. The disposition of the LZCC protein residues in wenxiang diagram provides new insights into the protein-protein interaction mechanism. J Theor Biol. 2011;284(1). doi:10.1016/j.jtbi.2011.06.006

Zhou W, Yang Y, Xia J, et al. NEK2 Induces Drug Resistance Mainly through Activation of Efflux Drug Pumps and Is Associated with Poor Prognosis in Myeloma and Other Cancers. Cancer Cell. 2013;23(1). doi:10.1016/j.ccr.2012.12.001

Zhou, J., Wang, M., Mao, A., Zhao, Y., Wang, L., Xu, Y., Jia, H., & Wang, L. (2021). Long noncoding RNA MALAT1 sponging miR-26a-5p to modulate Smad1 contributes to colorectal cancer progression by regulating autophagy. Carcinogenesis, 42(11), 1370–1379. https://doi.org/10.1093/carcin/bgab069

Zhou, W. X. (2008). Multifractal detrended cross-correlation analysis for two nonstationary signals. Physical Review E, 77(6), 066211.

Zhou, Y., Xu, S., Xia, H., Gao, Z., Huang, R., Tang, E., & Jiang, X. (2019). Long noncoding RNA FEZF1-AS1 in human cancers. Clinica Chimica Acta, 497, 20-26.

Zhu C, Shao P, Bao M, et al. MiR-154 inhibits prostate cancer cell proliferation by targeting CCND2. Urol Oncol Semin Orig Investig. 2014;32:31.e9-31.e16. https://doi.org/10.1016/j.urolonc.2012. 11.013

Zhu M, Zhao S. Candidate gene identification approach: Progress and challenges. Int J Biol Sci. 2007;3(7). doi:10.7150/ijbs.3.420

Zielezinski A, Vinga S, Almeida J, Karlowski WM. Alignment-free sequence comparison: benefits, applications, and tools. Genome Biol. 2017;18:186. https://doi.org/10.1186/s13059-017-1319-7

Appendix

LNCRDBCC: A Manually Curated Database of lncRNA related to Cervical Cancer

Long non-coding RNAs (lncRNAs) are molecules that can alter cancer development and progression. Several databases have aimed to collate and arrange the experimental data supporting lncRNA-cancer relationships. For example, Lnc2Cancer is a manually curated database that was first created by Ning in 2015 (Ning et al., 2016) and updated by Gao in 2018 (Gao et al., 2018). It contains comprehensive information on how lncRNAs regulate cancer through different mechanisms. Another database, CRlncRNA, was developed by Wang and focuses on the functional roles of cancer-related lncRNAs (Wang et al., 2018). It also provides information on the clinical and molecular characteristics of these lncRNAs.

Furthermore, LncRNADisease gives information about lncRNA-disease associations, along with the addition of transcriptional regulatory relationships and a confidence score for each association (Bao et al., 2018). These databases are useful tools for researchers and clinicians to understand and explore the roles of lncRNAs in cancer. However, none of them focus specifically on lncRNAs related to cervical cancer, which is a significant gap in the current knowledge base.

To address this, LNCRDBCC, a manually curated database of lncRNAs related to cervical cancer, was developed to provide information on differentially expressed lncRNAs in cervical cancer. It is implemented using MySQL and aims to support researchers who study cervical cancer and its relation to lncRNAs. Unlike other databases that cover a broader range of lncRNAs, LNCRDBCC focuses exclusively on lncRNAs associated with cervical cancer. This makes it more specific and relevant for cervical cancer research. LNCRDBCC can be accessed at http://sls.uohyd.ac.in/new/lncrdbcc/index2.html.

MATERIALS AND METHODS

Data Acquisition:

The transcription profiles of lncRNA from the TCGA-CESC and GTEx projects were used to identify differentially expressed lncRNAs. A total of 731 lncRNAs showed differential expression between the tumor and normal samples. Apart from this, the PubMed database was searched for all literature published until 30th August 2023 to identify lncRNAs related to cervical cancer using the *entrez_search* function of the rentrez R package.

In order to retrieve the information, keywords used were: "long non-coding RNA", "lncRNA", "long non-coding", "cervical", "cervix", "HPV", and "human papillomavirus". Entrez search results with 524 hits reporting lncRNAs associated with cervical cancer development, progression, diagnosis, or treatment (Figure 1).

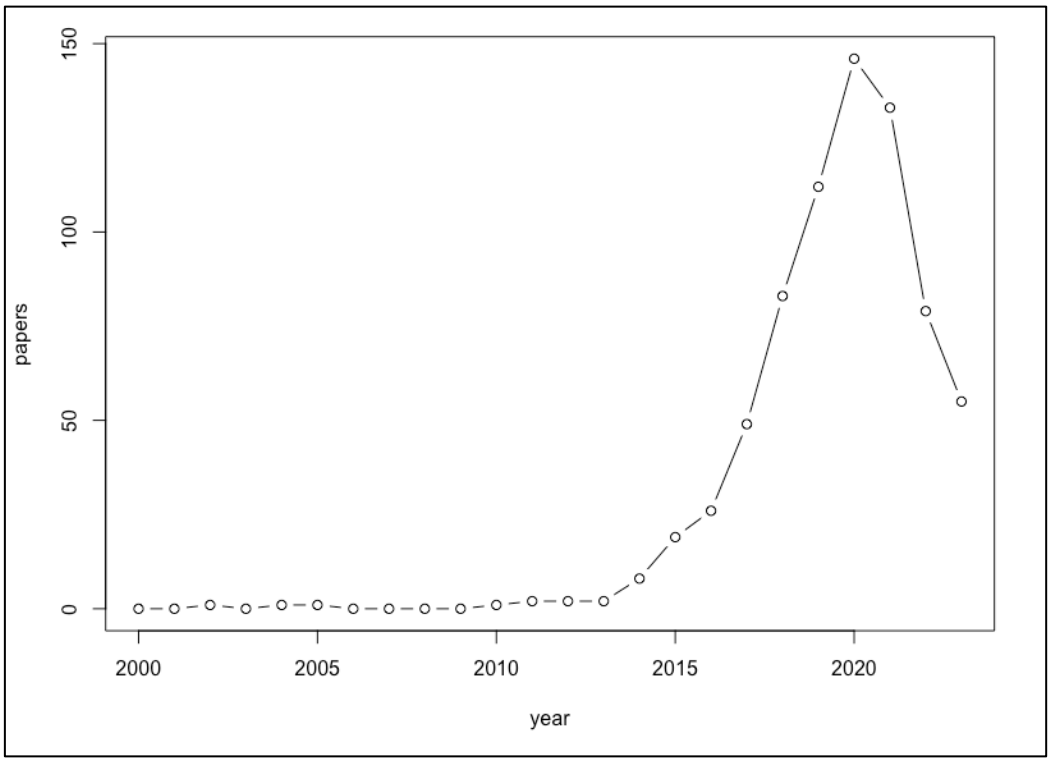


Figure 1: Number of research articles on lncRNA and cervical cancer

Several bioinformatics techniques, including survival, co-expression analysis, and functional enrichment, were used to further enhance our understanding of the biology behind differentially expressed lncRNAs.

DATABASE(LNCRDBCC) CONSTRUCTION:

The web graphical user interface (GUI) and the relational database, LNCRDBCC, were developed on the XAMPP platform (version 7.4.13). MySQL was used as the relational database management system (RDBMS). The front end was developed using HTML (Hypertext Markup Language), JavaScript (JS) and CSS (Cascading Style Sheets). PHP (Hypertext Preprocessor) scripting language enabled interaction between the back end and front end for query processing.

All the general information about lncRNAs was obtained from the HGNC (HUGO GENE NOMENCLATURE COMMITTEE) database (https://www.genenames.org/). The lncRNA sequences were downloaded from the Ensembl database (http://asia.ensembl.org/index.html). The Gene cards database (https://www.genecards.org/) was used to check whether the lncRNAs were associated with cervical cancer or other types of cancer. All the information was stored and managed using MySql data tables.

RESULTS

DATABASE FEATURES:

LNCRDBCC is a public database of long non-coding RNAs (lncRNAs) that have been manually curated. LNCRDBCC allows users to explore the regulatory functions of lncRNAs in cervical cancer development and progression. Currently, LNCRDBCC contains 731 entries of lncRNAs associated with cervical cancer.

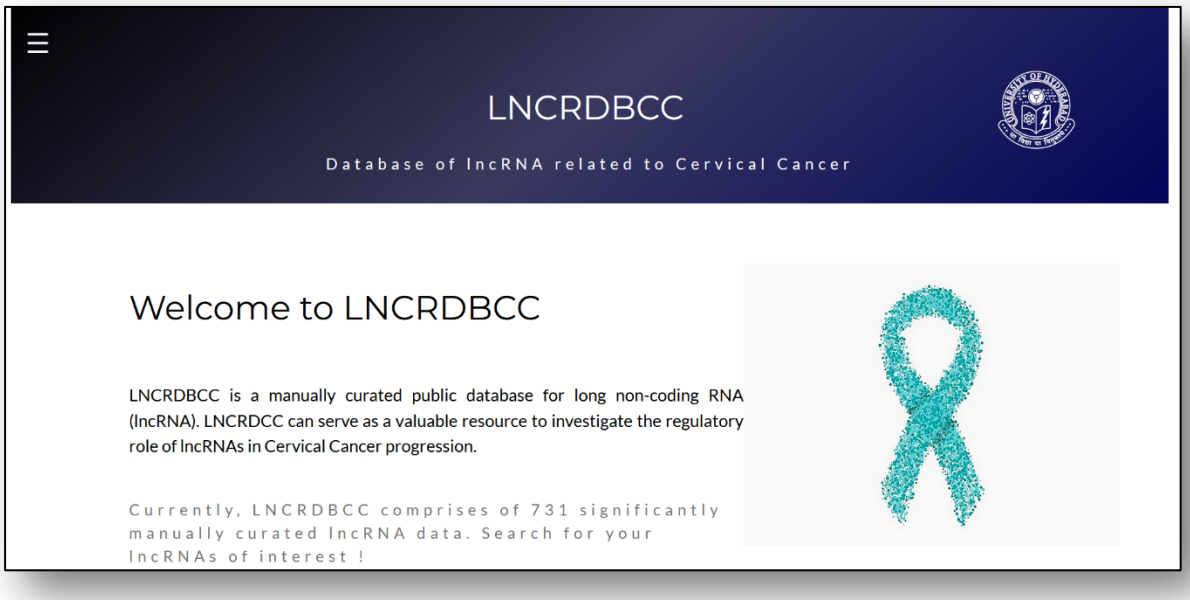


Figure 2: Home page of LNCRDBCC

The LNCRDBCC database contains lncRNAs that are involved in cervical cancer. Users can access the database through three main modules: the search module, the network module, and the download module. The search module allows users to query the database by various criteria, such as lncRNA name, gene symbol, or chromosomal location. The network module lets users visualize and analyze the interactions between lncRNAs and other molecules, such as proteins,

microRNAs, or DNA methylation sites. The download module allows users to download the entire database or a subset of it for offline analysis.

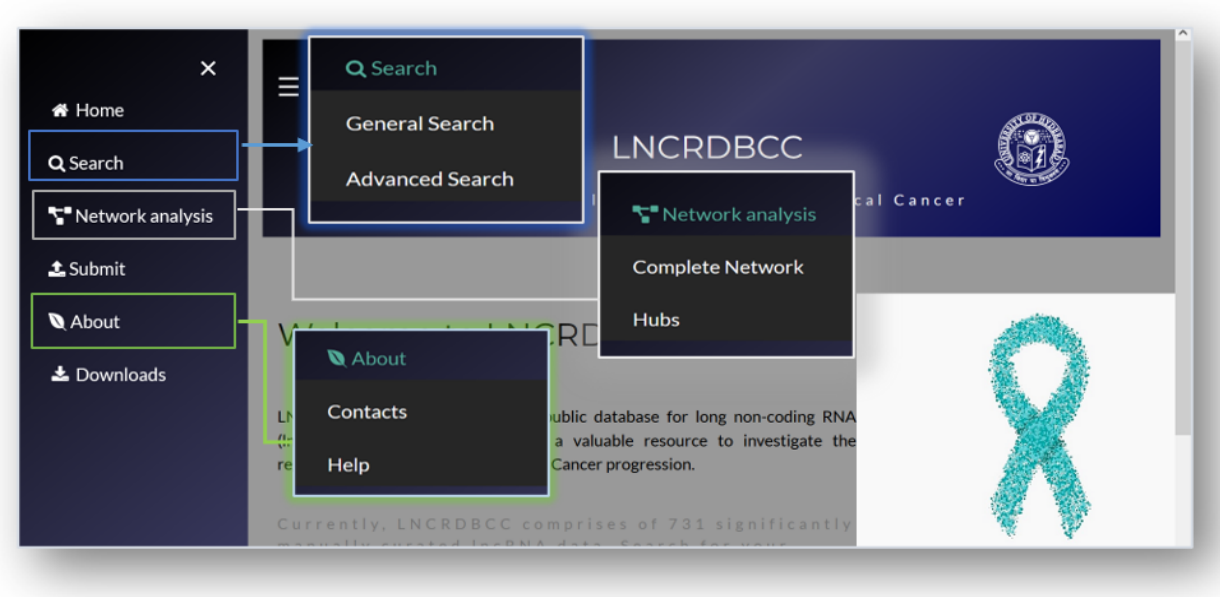


Figure 3: Snapshot of LNCRDBCC modules

SEARCH FUNCTIONALITY

The search module allows users to query the database by various criteria, either by 1) General search or 2) Advanced search

In "*General search*" users can search for lncRNA of interest in two alternative ways:

1) By lncRNA HGNC 'GENE symbol' or

2) By the sequence of lncRNA

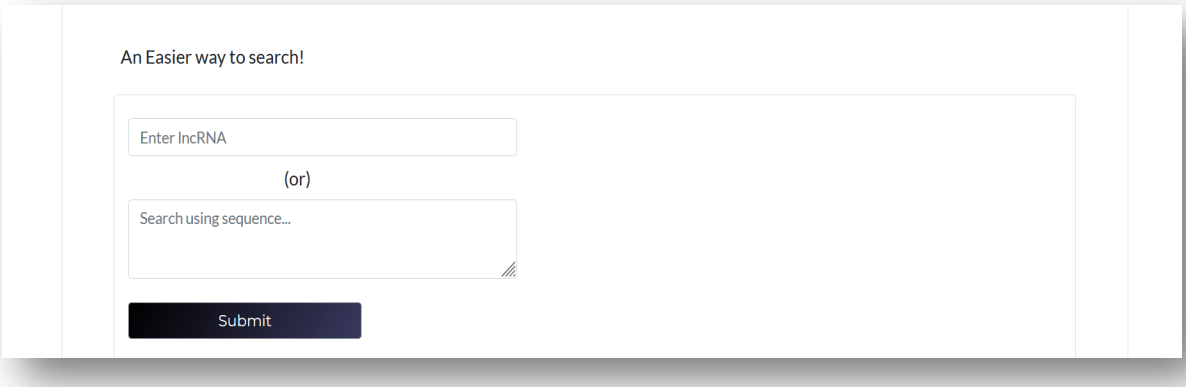


Figure 4: General Search module

The 'Advanced Search' interface allows users to apply filters that combine 'Expression Patterns' and 'Source' options for more specific queries. For instance, users can select 'Upregulated' and 'Literature' to find lncRNAs that have increased expression and are mentioned in the literature.

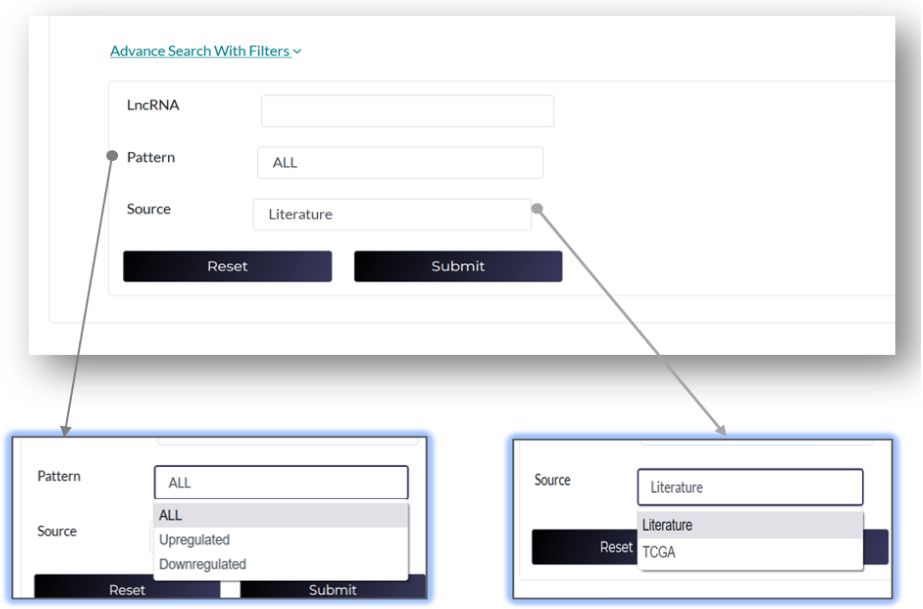
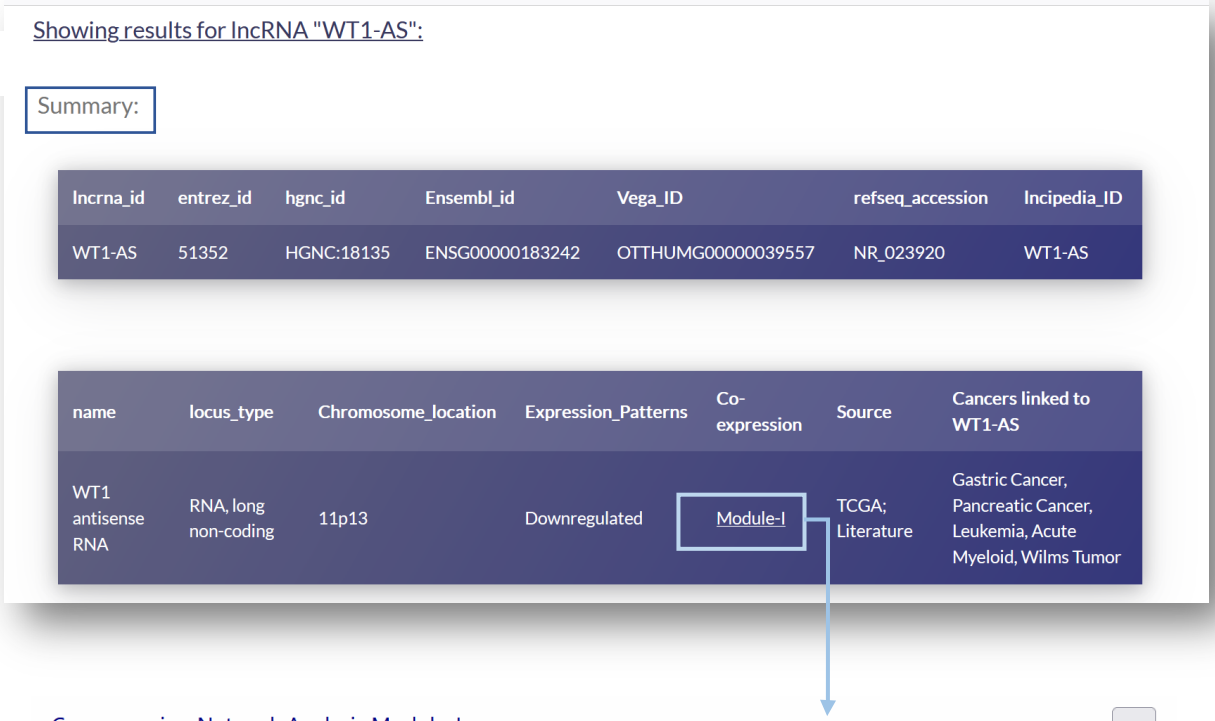


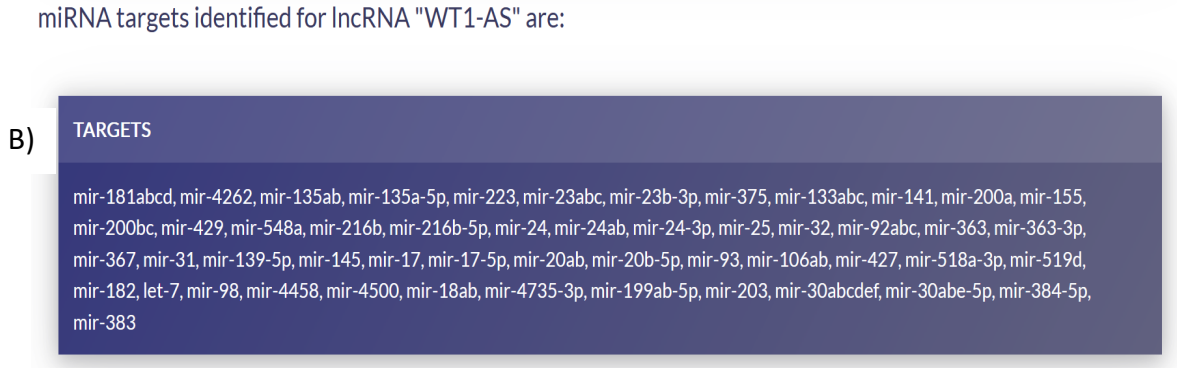
Figure 5: Snapshot of Advanced Search

SEARCH RESULTS

The search functionality provides a summary of the lncRNA of interest, such as WT1-AS, which is a hub lncRNA. The summary contains general information from the HGNC database, such as gene symbol, Entrez ID, Ensembl ID, Vega ID, RefSeq accession, Lncipedia ID, locus type and Chromosome location. It also shows the expression pattern from DEG analysis, the co-expression status from WGCNA analysis, and the source and related cancers from the Genecards database. The figure below illustrates the summary page for WT1-AS.

A)



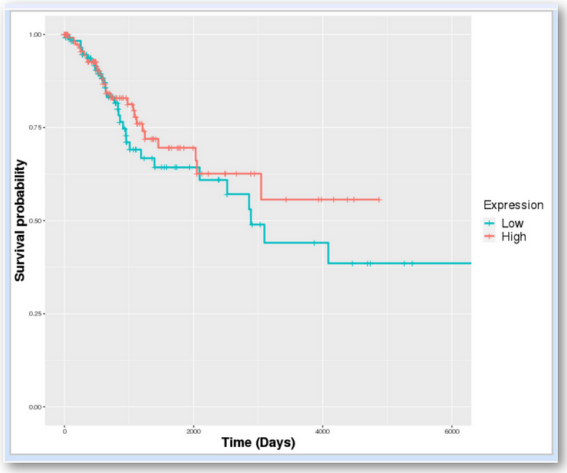


B)

Figure 6: Snapshots of results page showing A)Summary B) miRNA targets for lncRNA of interest

SURVIVAL PLOT

A)

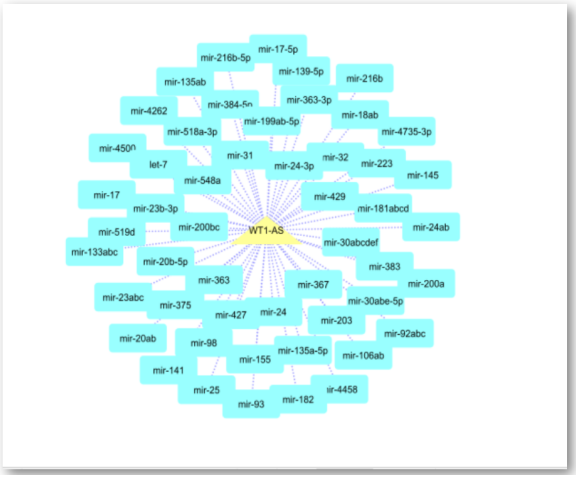


Kaplan-Meier curve (p-value)

0.339117

LncRNA-Target NETWORK

B)



RNA SEQUENCE

C)

Sequence

>WT1-AS
AGGGGCCGGCGGCGCGGGGAAGAGGAGGAGCCAGGAGGCTCGGCCGCTCCATTCACTCAG
CAGCCCAAGCCCGGCCAGGCAGCGCGCGCTGCTCCTTGGGGGCGCCCTCGCAGCTGGGGT
GAGGGCGGGTCCGAGGGAGCGTCCCTCTCGGAGACACCCTCCTCTTCAACCCAACACGCG
CTCTCAGCTGGCCGGTGTGGCACAACCCTCCCCCCACCTCCCCTTCTTGGCCTTCGCCAG
TGTGGGAGCCTCGTCTCACTGGAGAGTCCGTTCAGGTAAGCAGTGAGTCCGGGCAGGGTC
GAATGCGGTGGGAGTAGAGATGGGGTTGTTAATCTAAAAGAGTGGTTTGGAGGGAGGGGC
AGGTAGGCTCCAAGAGGGCGGGAACTTCAGGAAAGCGCAGAAAGAACTCAGTCAAGCTTC
AAAGAAGCGTATACCCTTGCTTTGCACCCTTGAGTCTGGCTCTTGCTTCTAGAACTGTCG
GGTAGGGGAGTGGCCATGATCTTGGGGTCACTGGGGAAAGAGGTGGCCCTTTGTGGGTTC
CAGAGGTCGGGACTGAAAGCGAACGTCCCTCAGTTGGGTGGCCGGGAGTTGCACCAGTTA
TAAGCCGGGCGTATTTCGATCGGCTAGTAGTTGTTTACTTGGTGAGGTTTGCGCCAGGCG

Figure 7: Screenshots of results page showing A) Survival plot B) lncRNA- Target miRNA network C) FASTA Sequence

The lncRNA targets found in the analysis are also displayed on the results page (Figure 6B). Additionally, it displays a graphical representation of the interactions between lncRNA and miRNA, generated with the Cytoscape platform (Figure 7B). A Kaplan-Meier survival plot and the associated p-value for every lncRNA are also included on the results page (Figure 7A). The FASTA sequence of the relevant lncRNA is included in the RNA sequence section (Figure 7C).

DOWNLOADS MODULE

Through the downloads module, users can obtain comprehensive data from LNCRDBCC. Files in CSV format with general data, expected targets, lncRNA sequences, and network analysis properties tables are available for download.

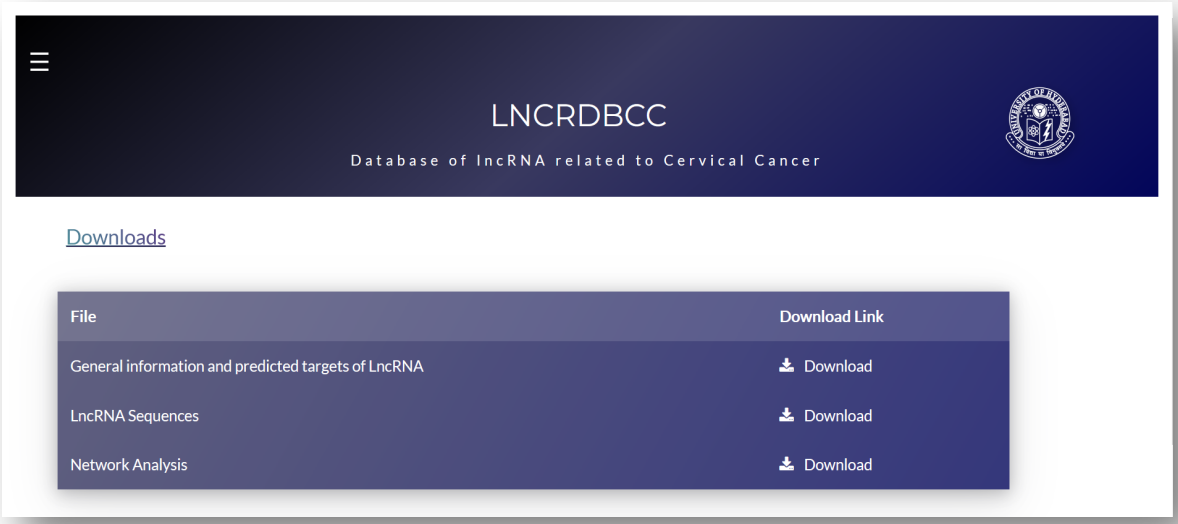


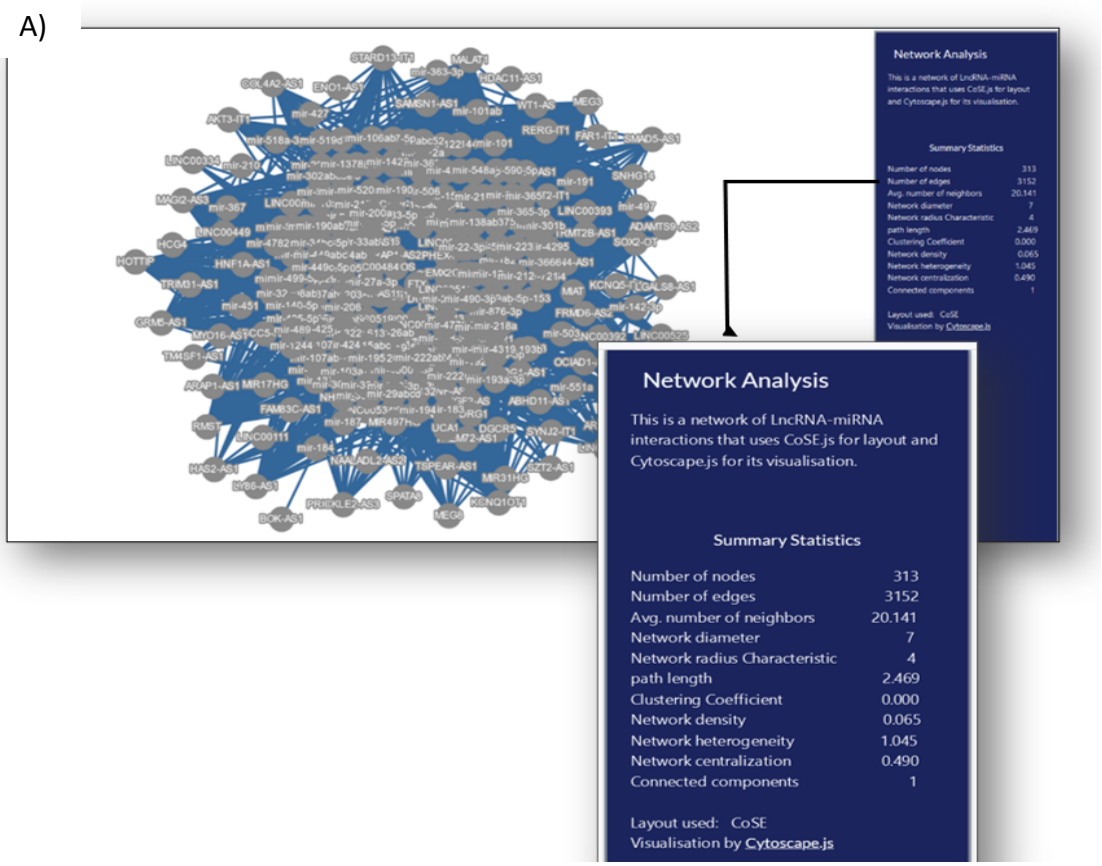
Figure 8: Snapshot of Downloads page

NETWORK ANALYSIS MODULE

LNCRDBCC includes two categories for the Network Analysis module: 'Complete network' and 'Hub network'. The 'Complete network' category displays an interactive network of 108 lncRNAs and their target miRNAs. The summary statistics and the network properties (such as Closeness Centrality, Degree, Eccentricity, Betweenness Centrality, Topological Coefficient and Average Shortest PathLength) are shown in Figure 9. In the 'hub network' section, we developed an interactive network visualization of the three novel hub lncRNAs and their target

miRNAs. We also added a table that lists the novel lncRNAs, the number of targets (miRNA) for each lncRNA, and the functions of the miRNA targets.

A)



NETWORK ANALYSIS

B)

name	ClosenessCentrality	Degree	Eccentricity	BetweennessCentrality	TopologicalCoefficient	AverageShortestPathLength
SNHG14	0.63803681	172	4	0.136679956	0.144518272	1.567307692
KCNQ1OT1	0.615384615	164	4	0.119638598	0.14880394	1.625
MEG3	0.564195298	142	4	0.096310409	0.157459319	1.772435897
MALAT1	0.500802568	113	4	0.047892109	0.170422407	1.996794872
ADAMTS9-AS2	0.474885845	97	4	0.038865478	0.159256873	2.105769231
MIAT	0.477794793	93	4	0.037185599	0.171199663	2.092948718
MAGI2-AS3	0.459499264	88	4	0.031348676	0.159090909	2.176282051
SOX2-OT	0.455474453	85	4	0.031613866	0.156946183	2.195512821
GRM5-AS1	0.451519537	82	4	0.034662351	0.158017644	2.21474359
HOTTIP	0.451519537	82	4	0.025842521	0.184613389	2.21474359

Figure 9: Snapshot of 'Complete Network' module A) lncRNA-target interactive network and B) Properties of lncRNA hubs

The **About module** consists of two sections: Contacts and Help. The Contacts page provides the contact details of the database administrators who can assist the users with any queries or issues. The Help page guides the users on how to navigate and use the database effectively.
The **Submit module** allows the users to contribute information related to lncRNAs by filling out a form. The submitted data will be reviewed and added to the database after verification.

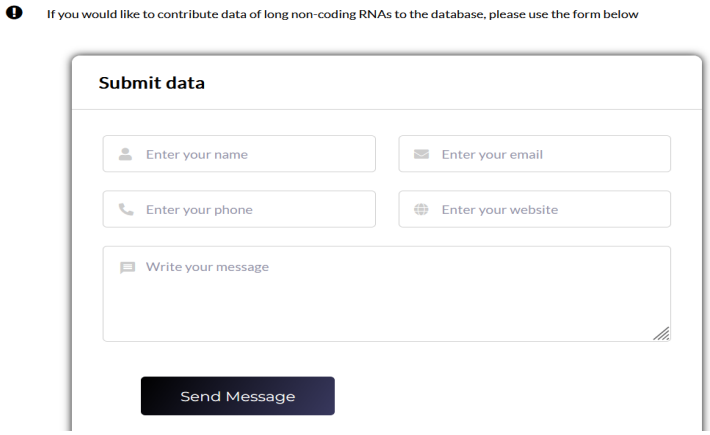


Figure 10: Snapshot of Submit Page

Conclusion

In order to investigate the regulatory function of long noncoding RNAs (lncRNAs) in cervical cancer, a network of differentially expressed lncRNAs and their target microRNAs was constructed. Nineteen hub lncRNAs involved in network regulation were identified, and it was also observed that some of these had also been described in other types of cancer. Three of the hub lncRNAs, however, were novel and had not been associated to any other cancers.
LNCRDBCC, a web-based database of human lncRNA data related to cervical cancer is developed to organize and present the data obtained from various bioinformatic analyses, such as DEG analysis, functional enrichment, coexpression analysis, survival analysis, and network analysis. The current version of LNCRDBCC contains 731 manually curated lncRNA entries. LNCRDBCC is an open resource that allows users to query and analyze the regulatory role of lncRNAs in cervical cancer, which may facilitate lncRNA research and the development of lncRNA-targeted therapeutics.

References

Bao, Z., Yang, Z., Huang, Z., Zhou, Y., Cui, Q., & Dong, D. (2019). LncRNADisease 2.0: an updated database of long non-coding RNA-associated diseases. *Nucleic acids research*, 47(D1), D1034-D1037.

Gao, Y., Wang, P., Wang, Y., Ma, X., Zhi, H., Zhou, D., ... & Li, X. (2019). Lnc2Cancer v2. 0: updated database of experimentally supported long non-coding RNAs in human cancers. *Nucleic acids research*, *47*(D1), D1028-D1033.

Ning, S., Zhang, J., Wang, P., Zhi, H., Wang, J., Liu, Y., ... & Li, X. (2016). Lnc2Cancer: a manually curated database of experimentally supported lncRNAs associated with various human cancers. *Nucleic acids research*, 44(D1), D980-D985.

Wang, J., Zhang, X., Chen, W., Li, J., & Liu, C. (2018). CRlncRNA: a manually curated database of cancer-related long non-coding RNAs with experimental proof of functions on clinicopathological and molecular features. *BMC medical genomics*, 11, 29-37.

Publications

Manuscripts published:

1. **Thippana M**, Dwivedi A, Das A, Palanisamy M, Vindal V. Identification of key molecular players and associated pathways in cervical squamous cell carcinoma progression through network analysis. Proteins. 2023; 1 - 15. doi:10.1002/prot.26502.
2. Thummadi NB, **Thippana M**, Vindal V, P. M. Prioritizing the candidate genes related to cervical cancer using the moment of inertia tensor. Proteins. 2021;1-9. doi:10.1002/prot.26226.

Manuscripts Communicated:

1. **Thippana M**, Thummadi NB, Vindal V and Manimaran P. Prioritizing cervical cancer candidate genes using chaos game and fractal-based time series approach. *Theory Biosci*. 2023 [revision submitted].

Manuscripts under preparation:

1. Integrative networks analysis to uncover regulatory elements associated with cervical cancer progression
2. Exploring lncRNA-mediated ceRNA regulatory mechanisms in cervical cancer via a systems biology approach.
3. Exploring HPV sample-specific prognostic players associated with cervical cancer

Presented in the following conferences

1. ISMB/ECCB 2023 (31st Conference on Intelligent Systems For Molecular Biology and 22nd Annual European Conference on Computational Biology), France, 2023.
2. GIWXXXI/ISCB-AsiaV2022(31stInternationalConferenceonGenomeInformatics and ISCB Asia V 2022), National Cheng-Kung University (NCKU), Taiwan, 2022.
3. ASCS2022 Asian Student Council Symposium 2022, December 10-11, 2022.
4. The 20th International Conference on Bioinformatics (InCoB 2021) held at Kunming, Yunnan, China (Virtual mode), November 2021.
5. Conference on “Proteomics in Agricultural and Healthcare" held at School of Life Sciences, University of Hyderabad, March 2021.
6. National workshop on Network Science (NetSci2020) held at School of Physics, University of Hyderabad, March 2020.
7. 1st TCGA Conference in India on “Multi-Omics Studies in Cancer: Learnings from The Cancer Genome Atlas (TCGA)” held at IISER-Pune, September 2019.

Received: 15 September 2022 | Revised: 23 March 2023 | Accepted: 6 April 2023

DOI: 10.1002/prot.26502

RESEARCH ARTICLE

PROTEINS WILEY

Identification of key molecular players and associated pathways in cervical squamous cell carcinoma progression through network analysis

Mallikarjuna Thippana[1] | Ayushi Dwivedi[1] | Abhishek Das[1] | Manimaran Palanisamy[2] | Vaibhav Vindal[1]

[1]Department of Biotechnology and Bioinformatics, School of Life Sciences, University of Hyderabad, Hyderabad, India

[2]School of Physics, University of Hyderabad, Hyderabad, India

Correspondence
Vaibhav Vindal, Department of Biotechnology and Bioinformatics, School of Life Sciences, University of Hyderabad, Gachibowli, Hyderabad 500046, India.
Email: vaibhav@uohyd.ac.in

Abstract

Cervical cancer is the primary cause of mortality among women in developing countries. Preventing cervical cancer is partially possible through early vaccination against the human papillomavirus, the most common cause of the disease. Nevertheless, it is imperative to understand the genetics of the disease progression to develop new therapeutic strategies. The present study aims to identify potential genes and associated pathways associated with cervical squamous cell carcinoma progression. We used an integrative approach by combining differential expression analysis, network biology, and functional enrichment analysis with survival analysis. In the present study, differential expression analysis of the microarray-based gene expression profiles of cervical cancer resulted in identifying a total of 544 significantly differentially expressed genes (DEGs). Further, centrality and network vulnerability analysis of the protein–protein interaction network (PPIN) and not well documented in cervical cancer, resulted in seven proteins (FN1, MCM5, TRIP13, KIF11, TTK, CDC45, and BUB1B), in which four proteins were vulnerable. These genes are mostly enriched in biological processes of cell division, mitotic nuclear division, cell cycle checkpoint, and cell proliferation in gene ontology analysis. The KEGG pathway enrichment analysis of the proteins lists them as mainly associated with the cell cycle. In the survival analysis, it was found that the genes MCM5, FN1, KIF11, and CDC45 were statistically significant prognostic factors for cervical cancer. The outcome of the current study identifies and explores the key role of the candidate genes involved in the progression of cervical cancer.

KEYWORDS
cervical cancer, differentially expressed genes, gene ontology, network vulnerability analysis, protein–protein interactions

1 | INTRODUCTION

Women worldwide suffer high morbidity and mortality rate because of cervical cancer (CC) which holds fourth rank among different types of cancers.[1,2] It is divided into two subtypes, that is, squamous cell carcinoma (80%–90%) and adenocarcinoma (10%–20%). India contributes at least one-fourth of the disease burden globally.[3] Human papillomavirus (HPV) is the primary risk factor for CC. However, weak immune system, smoking, birth control pills, multiple sexual partners, also play an imminent role as risk factors.[4] It is evident that tumor

Received: 14 June 2021 | Revised: 7 August 2021 | Accepted: 16 August 2021

DOI: 10.1002/prot.26226

RESEARCH ARTICLE

PROTEINS WILEY

Prioritizing the candidate genes related to cervical cancer using the moment of inertia tensor

Neelesh Babu Thummadi[1] | Mallikarjuna T.[2] | Vaibhav Vindal[2] | Manimaran P.[3]

[1]Department of Animal Biology, School of Life Sciences, University of Hyderabad, Gachibowli, Hyderabad, India

[2]Department of Biotechnology and Bioinformatics, School of Life Sciences, University of Hyderabad, Gachibowli, Hyderabad, India

[3]School of Physics, University of Hyderabad, Gachibowli, Hyderabad, India

Correspondence
Manimaran P., School of Physics, University of Hyderabad, Gachibowli, Hyderabad. Telangana 500046, India.
Email: manimaran@uohyd.ac.in

Funding information
DST-SERB GoI Project, Grant/Award Number: YSS/2015/000949; ICMR, India

Abstract

It is well known that cervical cancer poses the fourth most malignancy threat to women worldwide among all cancer types. There is a tremendous improvement in realizing the underlying molecular associations in cervical cancer. Several studies reported pieces of evidence for the involvement of various genes in the disease progression. However, with the ever-evolving bioinformatics tools, there has been an upsurge in predicting numerous genes responsible for cervical cancer progression and making it highly complex to target the genes for further evaluation. In this article, we prioritized the candidate genes based on the sequence similarity analysis with known cancer genes. For this purpose, we used the concept of the moment of inertia tensor, which reveals the similarities between the protein sequences more efficiently. Tensor for moment of inertia explores the similarity of the protein sequences based on the physicochemical properties of amino acids. From our analysis, we obtained 14 candidate cervical cancer genes, which are highly similar to known cervical cancer genes. Further, we analyzed the GO terms and prioritized these genes based on the number of hits with biological process, molecular functions, and their involvement in KEGG pathways. We also discussed the evidence-based involvement of the prioritized genes in other cancers and listed the available drugs for those genes.

KEYWORDS
candidate genes, cervical cancer, dendrogram, gene ontology, tensor analysis

1 | INTRODUCTION

With the drastic changes in the day-to-day lifestyle, the survival of humans is being hindered by the peril of noncommunicable diseases. Among these diseases, cancer might hold the first place to cause most deaths worldwide.[1] Despite the clinical advancements in cancer therapy, the mortality rate is still growing, and the reason for this imbalance rises from several facts like genetic variations among the individuals, mutational rate, substantial increase in the aged population, socio-economic variations, and so forth.[2] Cancer is a collective term for several related diseases. There are more than 100 varieties of cancers. The name of the cancer is termed based on the type of tissue or the cell of origin of cancer. Basically, in all cancers, the cells begin to divide uncontrollably and invade other tissues of the body, causing death. Among all cancers, cervical cancer is the fourth most malignancy threat for women worldwide. human papillomavirus (HPV) is the leading cause of most cervical cancer cases. There are several other risk factors like intercourse at an early age, multiple sexual partners, immunosuppression, and smoking.[3] Several vaccine programs have been shown beneficial in curbing malignancy.[4] Nevertheless, it is imperative to study the molecular mechanisms involved in cervical cancer. With advancements in molecular techniques, systems biology, bioinformatics, next-generation sequencing, microarray, and so forth, there has been a substantial improvement in understanding the underlying molecular mechanisms. High-throughput molecular techniques in synergy with computational analysis helped identify numerous cervical cancer genes, miRNAs, circRNAs, and lncRNAs.

TCGA

IISER PUNE

Prashanti Cancer Care Mission
Reaching out with excellence

PERSISTENT

Affordable Approaches to Cancer
A research challenge from India and the UK

CANCER RESEARCH UK

जैवप्रौद्योगिकी विभाग
Department of Biotechnology
Ministry of Science & Technology
Government of India

IndiaAlliance
DBT wellcome

1ST TCGA CONFERENCE AND WORKSHOP IN INDIA

Multi-Omics Studies in Cancer
Learnings from The Cancer Genome Atlas (TCGA)

Certificate

This certificate is awarded to **Mallikarjuna Thippanaa** *for participating in the Poster Presentation at the* ***"TCGA Conference"*** *held at IISER, Pune, India on 21 – 22 September, 2019*

JEAN C ZENKLUSEN	**DR C B KOPPIKER**	**L S SHASHIDHARA**	**DR ANAND DESHPANDE**
Director, TCGA, NCI, NIH	Medical Director, Prashanti Cancer Care Mission	Professor, IISER, Pune	Chairman and Managing Director, Persistent Systems

NetSci2020

प्रतिष्ठित संस्थान
INSTITUTION OF EMINENCE

National Workshop on Network Science
13-15 March, 2020
School of Physics, University of Hyderabad

This is to certify that T. Mallikarjuna

has ~~delivered an invited talk~~/presented work/participated

in the **NetSci2020**.

Dr. P. Manimaran
(Convener)

Dr. Ashok S Vudayagiri
(Convener)

हैदराबाद विश्वविद्यालय
University of Hyderabad

Proteomics Society, India (PSI)

प्रतिष्ठित संस्थान
INSTITUTION OF EMINENCE
राष्ट्रीय अपेक्षाएँ, वैश्विक मानक
National Needs, Global Standards

Virtual Conference on Proteomics in Agriculture and Healthcare

School of Life Sciences, University of Hyderabad

Certificate of Merit

This is to certify that Mr. Thippana Mallikarjuna S/o T Thirupaiah from University of Hyderabad, Hyderabad has participated in oral presentation in the conference held during March 13-14, 2021.

Prof. MV Jagannadham

Convener

Prof. S Rajagopal

Organising Secretary

Prof. S Dayananda

Dean, School of Life Sciences

InCoB 2021
Kunming · China

CERTIFICATE OF PARTICIPATION

This certificate is proudly presented to:

Mallikarjuna Thippana

For your participation in Poster Presentation during 20th International Conference on Bioinformatics (InCoB 2021)

(November 6 – 8, 2021)

Assoc. Prof. Dr. Mohammad Asif Khan

President, APBioNet

Prof. Dr. Yun Zheng

InCoB2021 Conference Chairman

Fostering the Growth of Bioinformatics in the Asia-Pacific

Certificate of Participation & Presentation

INTERNATIONAL SOCIETY FOR COMPUTATIONAL BIOLOGY

December 17, 2022

Mallikarjuna Thippana
S-67, Computational Functional Genomics Laboratory,
Dept. of Biotechnology and Bioinformatics,
School of Life Sciences,
University of Hyderabad,
Hyderabad - 500046,
India

Subject: CERTIFICATION OF PARTICIPATION AND PRESENTATION

On behalf of the Organizing Committee for Asian Student Council Symposium 2022 held Virtually, thank you for your participation and presentation of "*Exploring HPV sample-specific prognostic players associated with cervical cancer* "

This letter certifies that Mallikarjuna Thippana was a participant at Asian Student Council Symposium 2022 and presented the above research.

Asian Student Council Symposium 2022, the 1st such conference, was hosted by the International Society for Computational Biology (ISCB) December 10 - 11, 2022.

Please visit https://www.ascs2022.iscbsc.org/ for additional conference information.

Yours sincerely,

Diane E Kovats

Diane E. Kovats, CAE, CMP, DES
Chief Executive Officer
International Society for Computational Biology (ISCB)
e: dkovats@iscb.org
www.iscb.org

Certificate of Participation & Presentation

International Conference on Genome Informatics

Dear Mallikarjuna Thippana,

University of Hyderabad, Telangana, IndiaDear

Subject: CERTIFICATION OF PARTICIPATION AND PRESENTATION

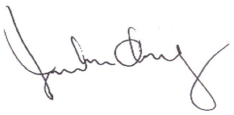
On behalf of the Conference Chairs for GIW XXXI/ISCB-Asia V conference held on site, thank you for your participation and presentation of "*Ascertainment of non-coding genes as key molecular players in cervical squamous cell carcinoma through the systems biology approach* "

This letter certifies that Mallikarjuna Thippana was a participant at GIW XXXI/ISCB-Asia V conference and presented the above research.

GIW XXXI/ISCB-Asia V conference, the 31st such conference, was hosted by National Cheng Kung University December 12 - 14, 2022.

Please visit https://www.iscb.org/giw-iscb-asia2022 for additional conference information.

Sincerely,



Conference Chairs
Jung-Hsien Chiang, National Cheng Kung University, Taiwan
Paul Horton, National Cheng Kung University, Taiwan
2022giw@gmail.com

Certificate of Participation & Presentation

INTERNATIONAL SOCIETY FOR COMPUTATIONAL BIOLOGY

July 27, 2023

Mallikarjuna Thippana
PhD Candidate
Computational and Functional Genomics laboratory
Dept. of Biotechnology and Bioinformatics
School of Life Sciences
University of Hyderabad
Hyderabad - 500046
India

Subject: CERTIFICATION OF PARTICIPATION AND PRESENTATION

On behalf of the Organizing Committee for Intelligent Systems for Molecular Biology/European Conference on Computational Biology (ISMB/ECCB) 2023 held in Lyon, France, thank you for your participation and presentation of "*Identification and characterization of key long non-coding RNAs involved in cervical cancer progression using systems biology approach* "

This letter certifies that Mallikarjuna Thippana was a participant at Intelligent Systems for Molecular Biology/European Conference on Computational Biology (ISMB/ECCB) 2023 and presented the above research.

Intelligent Systems for Molecular Biology/European Conference on Computational Biology (ISMB/ECCB) 2023, the 31th annual conference, was hosted by the International Society for Computational Biology (ISCB) July 23 - 27, 2023. The ISMB conference has grown to become the world's largest bioinformatics/computational biology conference.
Please visit https://www.iscb.org/ismbeccb2023 for additional conference information.

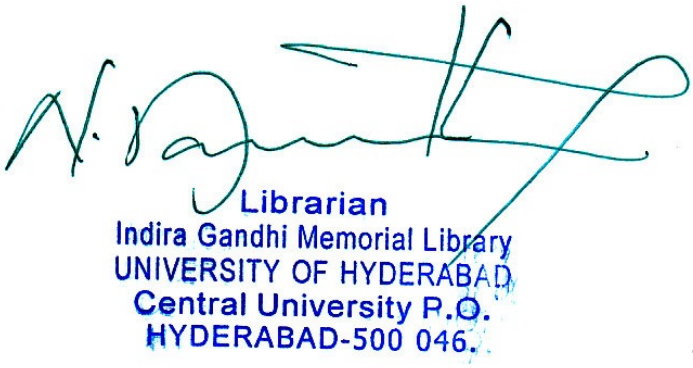
Yours sincerely,

Diane E Kovats

Diane E. Kovats, CAE, CMP, DES
Chief Executive Officer
International Society for Computational Biology (ISCB)
e: dkovats@iscb.org

"Integrative studies to explore key molecular players involved in cervical squamous cell carcinoma "

by Mallikarjuna Thippana



Submission date: 31-Jan-2024 02:32PM (UTC+0530)
Submission ID: 2282793271
File name: Thesis_Mallikarjun.pdf (12.77M)
Word count: 23408
Character count: 130269

"Integrative studies to explore key molecular players involved in cervical squamous cell carcinoma "

ORIGINALITY REPORT

36%	7%	36%	4%
SIMILARITY INDEX	INTERNET SOURCES	PUBLICATIONS	STUDENT PAPERS

PRIMARY SOURCES

1	Mallikarjuna Thippana, Ayushi Dwivedi, Abhishek Das, Manimaran Palanisamy, Vaibhav Vindal. "Identification of key molecular players and associated pathways in cervical squamous cell carcinoma progression through network analysis", Proteins: Structure, Function, and Bioinformatics, 2023 Publication	17%
2	N. B. Thummadi, T. Mallikarjuna, V. Vindal, P. Manimaran. "Prioritizing the candidate genes related to cervical cancer using the moment of inertia tensor", Proteins: Structure, Function, and Bioinformatics, 2021 Publication	11%
3	Submitted to University of Hyderabad, Hyderabad Student Paper	2%
4	Neelesh Babu Thummadi, Mallikarjuna T., Vaibhav Vindal, Manimaran P.. "Prioritizing the candidate genes related to cervical cancer	2%

using the moment of inertia tensor", Proteins: Structure, Function, and Bioinformatics, 2021
Publication

5	N.B. Thummadi, S. Charutha, Mayukha Pal, P. Manimaran. "Multifractal and cross-correlation analysis on mitochondrial genome sequences using chaos game representation", Mitochondrion, 2021 Publication	<1 %
6	unsworks.unsw.edu.au Internet Source	<1 %
7	Submitted to 9561 Student Paper	<1 %
8	Mayukha Pal, B. Satish, K. Srinivas, P. Madhusudana Rao, P. Manimaran. "Multifractal detrended cross-correlation analysis of coding and non-coding DNA sequences through chaos-game representation", Physica A: Statistical Mechanics and its Applications, 2015 Publication	<1 %
9	www.frontiersin.org Internet Source	<1 %
10	www.ncbi.nlm.nih.gov Internet Source	<1 %
11	"Genetic Polymorphism and cancer susceptibility", Springer Science and Business	<1 %

Media LLC, 2021
Publication

12	Submitted to University College London Student Paper	<1%
13	Lin Liu, Zipeng Yu, Yang Xu, Cheng Guo et al. "Function identification of MdTIR1 in apple root growth benefited from the predicted MdPPI network", Journal of Integrative Plant Biology, 2020 Publication	<1%
14	Zhu-Hong You. "A semi-supervised learning approach to predict synthetic genetic interactions by combining functional and topological properties of functional gene network", BMC Bioinformatics, 2010 Publication	<1%
15	docplayer.net Internet Source	<1%
16	"Abstracts", Journal of Investigative Dermatology, 03/2008 Publication	<1%
17	Advances in Experimental Medicine and Biology, 2014. Publication	<1%
18	Submitted to City University Student Paper	<1%

res.mdpi.com

19	Internet Source	<1%
20	www.science.gov Internet Source	<1%
21	www.thno.org Internet Source	<1%
22	Lijun Bai, Qing Chen, Leiyu Jiang, Yuanxiu Lin, Yuntian Ye, Peng Liu, Xiaorong Wang, Haoru Tang. "Comparative transcriptome analysis uncovers the regulatory functions of long noncoding RNAs in fruit development and color changes of Fragaria pentaphylla", Horticulture Research, 2019 Publication	<1%
23	Sara E. Lipshutz, Clara R. Howell, Aaron M. Buechlein, Douglas B. Rusch, Kimberly A. Rosvall, Elizabeth P. Derryberry. "How thermal challenges change gene regulation in the songbird brain and gonad: implications for sexual selection in our changing world", Molecular Ecology, 2022 Publication	<1%
24	Shrestha Reshies, Min-Min Yu. "Expressions of Long Non-Coding RNAs in Carcinogenesis of Cervix: A Review", Open Journal of Obstetrics and Gynecology, 2018 Publication	<1%

25	Zhicheng Yan, Junyan Shi, Shuzhi Yuan, Dongying xu, Shufang Zheng, Lipu Gao, Caie Wu, Jinhua Zuo, Qing Wang. "Whole-transcriptome RNA sequencing highlights the molecular mechanisms associated with the maintenance of postharvest quality in broccoli by red LED irradiation", Postharvest Biology and Technology, 2022 Publication	<1 %
26	www.medrxiv.org Internet Source	<1 %
27	Submitted to Birla Institute of Technology and Science Pilani Student Paper	<1 %
28	Submitted to October University for Modern Sciences and Arts (MSA) Student Paper	<1 %
29	Pal, Mayukha, V. Satya Kiran, P. Madhusudana Rao, and P. Manimaran. "Multifractal detrended cross-correlation analysis of genome sequences using chaos-game representation", Physica A Statistical Mechanics and its Applications, 2016. Publication	<1 %
30	S Yilmaz. "Gene-Disease Relationship Discovery based on Model-driven Data	<1 %

Integration and Database View Definition", Bioinformatics, 11/27/2008
Publication

31	addi.ehu.es Internet Source	<1%
32	issx.confex.com Internet Source	<1%
33	www.aristeosegura.com.mx Internet Source	<1%
34	www.researchgate.net Internet Source	<1%
35	Submitted to Jawaharlal Nehru Technological University Student Paper	<1%
36	core.ac.uk Internet Source	<1%
37	Kai Liu, Shaoxi Chen, Ruoyi Lu. "Identification of important genes related to ferroptosis and hypoxia in acute myocardial infarction based on WGCNA", Bioengineered, 2021 Publication	<1%
38	Kim, Hee, Dae Lee, Ga Yim, Eun Nam, Sunghoon Kim, Sang Kim, and Young Kim. "Long non-coding RNA HOTAIR is associated with human cervical cancer progression", International Journal of Oncology, 2014. Publication	<1%

39	Matthew J. Kesic, Megan Meyer, Rebecca Bauer, Ilona Jaspers. "Exposure to Ozone Modulates Human Airway Protease/Antiprotease Balance Contributing to Increased Influenza A Infection", PLoS ONE, 2012 Publication	<1 %
40	Submitted to Grand Canyon University Student Paper	<1 %
41	Yeqing Mao, Chao Wen, Zitong Yang. "Construction of a Co-Expression Network for lncRNAs and mRNAs Related to Urothelial Carcinoma of the Bladder Progression", Frontiers in Oncology, 2022 Publication	<1 %
42	export.arxiv.org Internet Source	<1 %
43	qmro.qmul.ac.uk Internet Source	<1 %
44	www.biorxiv.org Internet Source	<1 %

Exclude quotes	On	Exclude matches	< 14 words
Exclude bibliography	On		

University of Hyderabad

(A central University established in 1974 by Act of Parliament)

(P.O.) Central University, Gachibowli

Hyderabad-500046

प्रतिष्ठित संस्थान
INSTITUTION OF EMINENCE

PLAGIARISM FREE CERTIFICATE

This is to certify that the thesis entitled **"Integrative studies to explore key molecular players involved in cervical squamous cell carcinoma"** submitted by **Mr. Thippana Mallikarjuna** bearing **Reg No: 18LTPH02**, in partial fulfillment of the requirements for the award of Doctor of Philosophy in the Department of Biotechnology and Bioinformatics, School of Life Sciences, is free from Plagiarism and has not been submitted previously in part or in full to this or any other University or Institution for the award of any degree or diploma.

The similarity index of this thesis as checked by the library of the University of Hyderabad is 36%. Out of this 30% similarity has been found to be identified from the candidate's own publication(s) which forms the major part of the thesis. The details of the student's publication are as follows:

Published in the following publications: List out the Publications here…..

1. **Thippana, M**, Dwivedi, A, Das, A, Palanisamy, M, Vindal, V. Identification of key molecular players and associated pathways in cervical squamous cell carcinoma progression through network analysis. Proteins. 2023; 91(8): 1173-1187. doi:10.1002/prot.26502

2. Thummadi NB, **Thippana M**, Vindal V, P. M. Prioritizing the candidate genes related to cervical cancer using the moment of inertia tensor. Proteins. 2021;1-9. doi:10.1002/prot.26226.

About 6% similarity was identified from the external sources in the present thesis which is according to the prescribed regulations of the university. All the publications related to the thesis have been appended at the end of the thesis. Hence the present thesis is considered to be plagiarism-free.

Dr. Vaibhav Vindal

[Supervisor]

Dr. VAIBHAV VINDAL
Associate Professor
Dept. of Biotechnology & Bioinformatics
School of Life Sciences
University of Hyderabad
Gachibowli, Hyderabad-500 046.

Prof P. Manimaran

[Co-supervisor]

Dr. P. Manimaran
Professor
School of Physics
University of Hyderabad
Hyderabad-500 046 (TS) India