

IDENTIFICATION OF DIFFERENTIALLY EXPRESSED GENES KEY GENES ASSOCIATED WITH ADENOCARCINOMA

Submitted in partial fulfillment of the requirements for the award of the degree of
Bachelor of Technology degree in
Bioinformatics

By

RESHMA.K (37220015)



**DEPARTMENT OF BIOINFORMATICS
SCHOOL OF BIO AND CHEMICAL ENGINEERING**

**SATHYABAMA UNIVERSITY
JEPPIAAR NAGAR, RAJIV GANDHI SALAI
CHENNAI - 600 119 TAMILNADU**

APRIL – 2021



SATHYABAMA

**INSTITUTE OF SCIENCE AND TECHNOLOGY
(DEEMED TO BE UNIVERSITY)**

Accredited with Grade "A" by NAAC

Jeppiaar Nagar, Rajiv Gandhi Salai, Chennai - 600 119, Tamil Nadu, India.



DEPARTMENT OF BIOINFORMATICS

BONAFIDE CERTIFICATE

This is to certify that this Project Report is the bonafide work of **RESHMA.K** (37220015) who carried out the project entitled "**IDENTIFICATION OF DIFFERENTIALLY EXPRESSED GENES KEY GENES ASSOCIATED WITH ADENOCARCINOMA**" under my supervision from September 2020 to April 2021

Internal Guide

DR. H JEMMY CHRISTY

Head of the Department

Submitted for Viva voce Examination held on

Internal Examiner

External Examiner

DECLARATION

I **RESHMA.K** (37220015) hereby declare that the Project Report entitled **“IDENTIFICATION OF DIFFERENTIALLY EXPRESSED GENES KEY GENES ASSOCIATED WITH ADENOCARCINOMA ”** done by me under the guidance of **Dr. H Jemmy Christy** (Internal guide) **Department of Bioinformatics at Sathyabama Institute of Science and Technology, Chennai, Tamil Nadu** is submitted in partial fulfillment of the requirements for the award of Bachelor of Technology degree in **BIOINFORMATICS**.

DATE:

PLACE:

SIGNATURE OF THE CANDIDATE

ACKNOWLEDGEMENT

I am pleased to acknowledge my sincere thanks to the Board of Management of **SATHYABAMA** for their kind encouragement in doing this project and for completing it successfully. I am grateful to them.

I convey my thanks to **Dr. H JEMMY CHRISTY** Head of the Department, Dept. of Bioinformatics for providing me necessary support and details at the right time during the progressive reviews.

I would like to express my sincere and deep sense of gratitude to my Project Guide **Dr. H JEMMY CHRISTY** for her valuable guidance, suggestions and constant encouragement paved way for the successful completion of my project work.

I wish to express my thanks to all Teaching and Non-teaching staff members of the Department of Bioinformatics who were helpful in many ways for the completion of the project.

RESHMA.K

TABLE OF CONTENT

CHAPTER NO	TITLE	PAGE NO
	ABSTRACT	i
	LIST OF ABBREVIATION	ii
	LIST OF TABLES	iii
	LIST OF FIGURES	iv
1	INTRODUCTION	
	1.1 OUTLINE OF THE PROJECT	
	1.1.1 CANCER	1
	1.1.2 ADENOCARCINOMA	2
	1.1.3 TYPES OF ADENOCARCINOMA	3
	1.1.4 INCIDENCE WORLD WIDE	11
	1.1.5 DIFFERENTIAL EXPRESSED ANALYSIS	13
	1.1.6 HUB GENE ASSESSMENT	15
	2. REVIEW OF LITERATURE	17
	3. AIM & OBJECTIVE	19
	3.1 AIM	
	3.2 OBJECTIVE	
	3.3 SCOPE OF THE WORK	
	3.3.1 MATERIALS AND METHODS	21
	3.1.2 TOOLS AND DATABASES USED	22
	3.1.3 WORK FLOW	
	3.1.4 METHODOLOGY	
	4. RESULTS AND DISCUSSION	32
	5. SUMMARY AND CONCLUSION	42
	REFERENCES	44

ABSTRACT OF THE WORK

Adenocarcinoma are malignant lumps which are formed from glandular structures in epithelial tissue. They occur from a variety of tissues. These types of cancer account for 80% to 95%. Malignant transformation of adenocarcinoma has been described in the literature as occurring mainly in relationship with the lung, esophageal, colorectal, pancreatic, prostate and gastric. Adenocarcinoma is a distinct type of lung, esophageal, gastric cancer with different clinicopathological and prognostic features. Differential gene expression analysis of adenocarcinoma associated lung cancer, esophageal cancer, prostate cancer, pancreatic cancer, colorectal cancer, gastric cancer data set which revealed the common genes, and those genes were further assessed for their hubness to screen the promising biomarker which could also serve as a therapeutic target. Natural compounds from marine resources are analysed for their affinity towards the proposed therapeutic targets.

.

Keywords: LUAD, Differentially expressed genes,adenocarcinoma,hubness, marine

LIST OF ABBREVIATION

GEO	-	Gene Expression Omnibus
DEG	-	Differentially Expressed Genes
FC	-	Fold Change
BioGRID	-	Biological General Repository for Interaction Datasets
LUAD	-	Lung adenocarcinoma
ADME	-	Absorption, Distribution, Metabolism and Excretion
BBB	-	Blood-Brain Barrier

LIST OF TABLES

TABLE NO	TITLE	PAGE NO
1.1	Statistical significances criteria	16
1.2	Parameters used for DEG screening	16
4.2	GEO dataset associated with adenocarcinoma	25
	DEG analysis	
5.1	Dataset associated with adenocarcinoma DEG analysis	31
5.2	Criterion of statistical significant upregulated DEGs	32
5.3	Upregulated and downregulated genes for significant dataset	33
5.4	Docked complex of 6QLU with compound available red algae	39
5.5	Interaction table of red algae	39
5.6	Docked complex of 6QLU with compound available green algae	40
5.7	Interaction table of green algae	40

LIST OF FIGURES

TABLE NO	TITILE	PAGE NO
1.1	Cancer	2
1.2	Adenocarcinoma	3
1.3	Lung cancer	5
1.4	Esophageal cancer	6
1.5	Pancreatic cancer	8
1.5.1	Stages of pancreatic cancer	8
1.6	Prostate cancer	10
1.7	Colorectal cancer	11
1.8	Gastric cancer	13
4.1	Proposed work plan	21
4.2	Steps involved in docking	28
5.3	Biogrid based network	30
5.4	Hub genes of GAL3	34
5.5	Admet analysis of red algae	40
5.6	Admet analysis of green algae	41

CHAPTER-1

INTRODUCTION

1.1 CANCER:

Cancer is a generic term used to describe a bunch of a minimum of 100 diseases that occur once malignant types of abnormal cell growth develop in one or additional body organs. Cancer arises once a series of genetic mutations remove the normal checks on cell growth. These cancer cells still divide and grow to provide tumours. Cancer cells will invade adjacent structures and unfold via the lymph or blood to distant organs.(Warren Kaplan.,2004)

Cancer happens by a series of sequent mutations in genes therefore that these mutations modification cell functions. Chemical compounds have a comprehensible role of forming sequence mutations and cancer cells. additionally, smoking involves many malignant neoplastic disease chemical compounds that result in carcinoma. The environmental chemical substances with malignant neoplastic disease properties influence directly or indirectly the living substance and nucleus of cells, and result in genetic disorders and gene mutations(Seyyed Hossein Hassanpour et al.,2017)

Differentiating options of malignant and benign lesion are well established; these embrace fast growth, exaggerated cell turn-over, invasive growth, metastases, tube or body fluid channel invasion for malignant lesions. There are several exceptions to those attributes of cancer. There are overlaps between benign and malignant lesions. Benign (non-malignant) tumors do show chromosome aberrations; body structure melanomas and blue nevi share mutations in G macromolecule. (Halliday A Idikio.,2011)

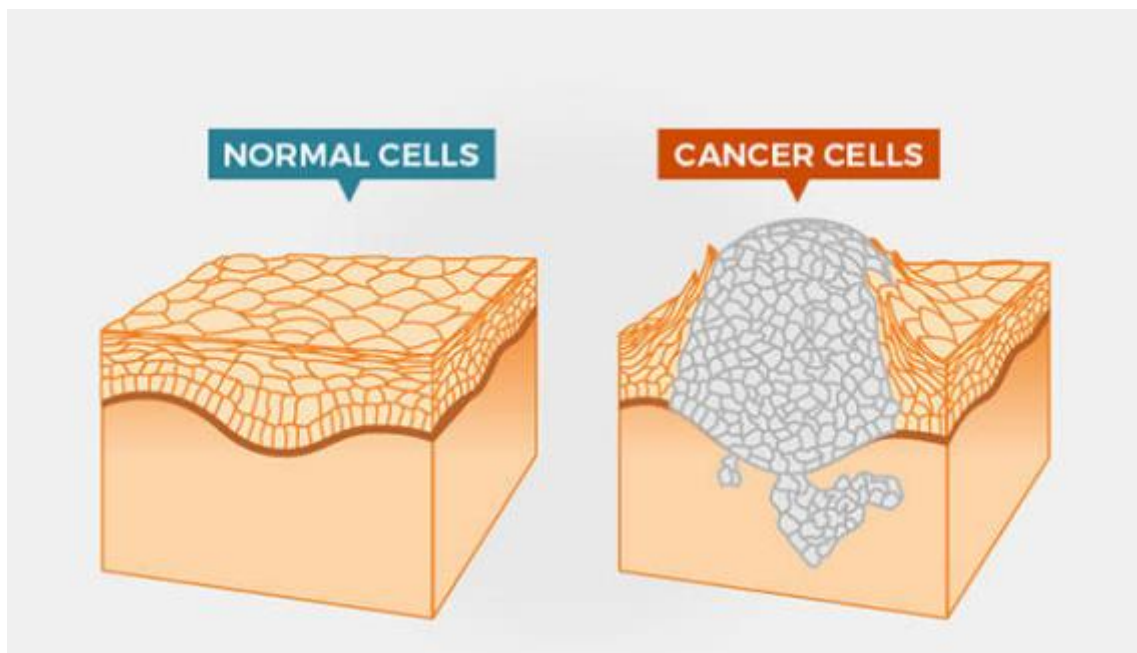


Fig 1.1 Cancer

(<https://www.cancer.gov/about-cancer/understanding/what-is-cancer>)

1.2 ADENOCARCINOMA:

Cancer that begins in secretory cells. Secretory cells are found in tissue that lines bound internal organs and makes and releases substances within the body, like mucous secretion, biological process juices, or alternative fluids. Most cancers of the breast, pancreas, lung, prostate, and colon are adenocarcinomas. (<https://www.cancer.gov/about-cancer/understanding/what-is-cancer>)

Adenocarcinoma is a kind of cancer that forms in mucus-secreting gland cells, that are found in tissues that line internal organs. When adenocarcinoma spreads from the initial site, it's represented as metastatic (Julie Marks.,2019).

Adenocarcinoma is the main histologic type, which accounts for half of all the cases.

For invasive adenocarcinoma, the new classification has introduced histological subtyping to the predominant pattern of growth of the neoplastic cells. In addition, four variants of invasive adenocarcinoma are recognized, namely invasive mucinous colloid, fetal, and enteric. The three variants that were considered in the previous classification have been eliminated, specifically mucinous cyst-adenocarcinoma, signet ring cell, and clear cell adenocarcinoma(E Kuhn et al 2018)

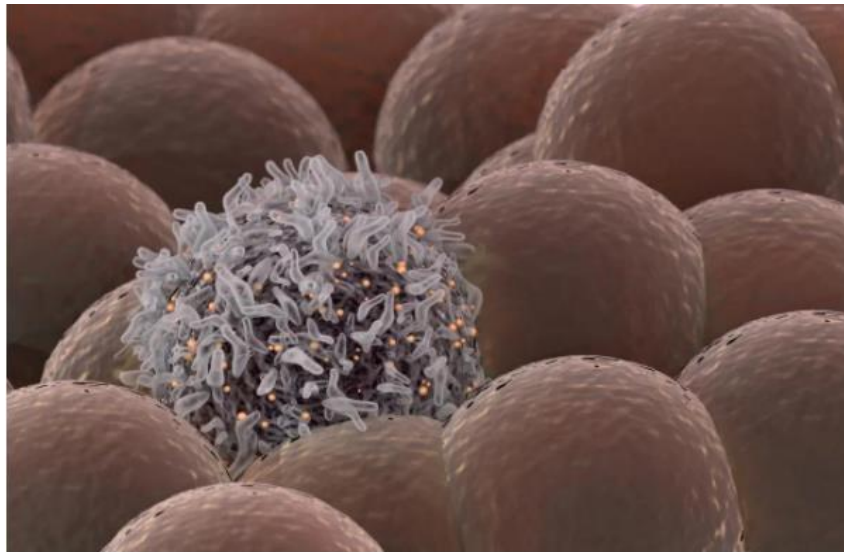


Fig1.2 Adenocarcinoma

(<https://www.medicalnewstoday.com/articles/318734#types>)

1.3 TYPES OF ADENOCARCINOMA

Adenocarcinomas begin in glands but can spread to other types of tissue and areas of the body. Most cancers in the following locations are adenocarcinomas:

1.3.1 LUNG CANCER

Lung cancer is the leading reason behind cancer-related deaths worldwide in both men and women. Lung cancer is classified based on two main histological groups: small cell respiratory organ cancer (SCLC, 15% of respiratory organ cancers) and non-SCLC (NSCLC, 85% of respiratory organ cancers). NSCLCs are usually subcategorized into carcinoma, squamous cell cancer (SqCC), and large cell cancer. Accumulating proof suggests that carcinoma represents a bunch of histologically and molecularly heterogeneous diseases even within the same histological subtype.(Kentarō Inamura .,2017)

Lung cancer, an extremely invasive, rapidly metastasizing and prevalent cancer which occur in both men and women. The individual status to tobacco-induced carcinoma is also keen about competitive gene–enzyme interactions that have an effect on activation or detoxification of pro-carcinogens and levels of deoxyribonucleic acid adduct formation also as determined by the integrity of endogenous mechanisms for repairing lesions in deoxyribonucleic acid. carcinoma is very heterogeneous which will arise in many alternative sites within the cartilaginous tube tree, so presenting extremely variable symptoms and signs betting on its anatomic location.(Hassan Lemjabbar-Alaoui et al., 2016)

The most common kind of lung cancer is adenocarcinoma; it contains around 40% of lung cancer. It arises from tiny airway epithelial, kind II alveolar cells, that secrete mucous secretion and different substances . Adenocarcinoma is the most common kind of lung cancer in smokers and nonsmokers in men and women regardless of their age. It tends to occur within the periphery of the respiratory organ , which could result in the addition of filters in cigarettes preventing massive particles from entering the lungs. This leads to deeper inhalation of tobacco smoke, resulting in peripheral lesions. Compared to different kinds of adenocarcinoma, adenocarcinoma tends to grow slower and encompasses a larger probability of being found before it's unfold outside of the lungs.(Cecilia Zappa et al., 2016)

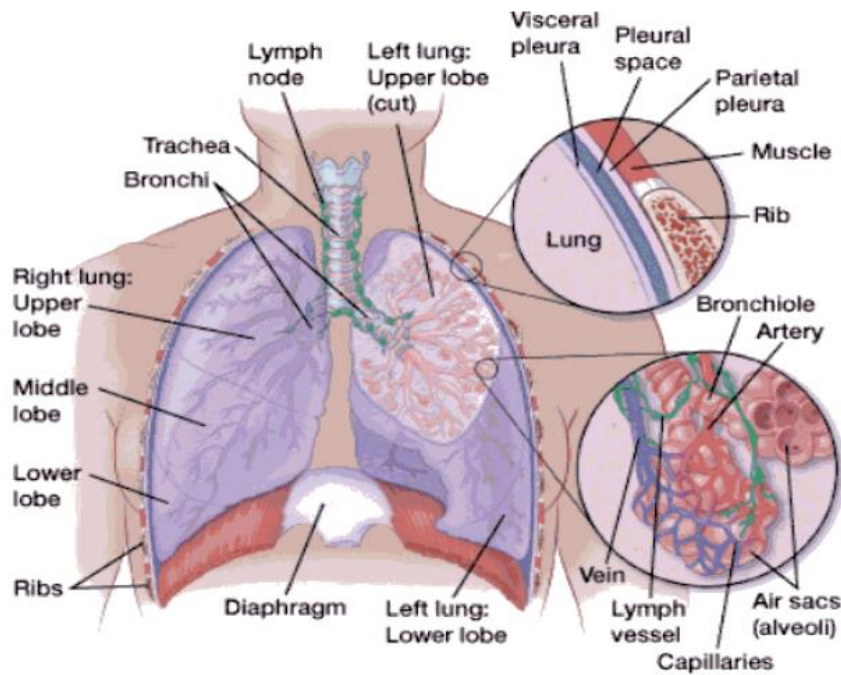


Fig 1.3 Lung cancer

(<https://www.cancer.org/content/dam/CRC/PDF/Public/8703.00.pdf>)

1.3.2 ESOPHAGEAL:

Esophageal cancer is a male-dominant aggressive malignancy and a number one reason for cancer-related mortality worldwide. squamous cell malignant neoplastic disease and carcinoma are the two predominant microscopic anatomy subtypes with varied geographical and racial distribution. Globally, squamous cell malignant neoplastic disease remains the most common histological sort. However, carcinoma has become the leading histological subtype, corresponding to an increase within the incidence of obesity, gastro-esophageal reflux disease and Barrett's muscle system.(Ghulam Abbas et al., 2017)

Adenocarcinoma of the esophagus occurs within the distal esophagus more or less simple fraction of the time and features a distinct link to reflux disease (GERD). Untreated GERD will accomplish Barrett's esophagus (BE), wherever the stratified

squamous epithelial tissue that unremarkably lines the oesophagus is replaced by a columnar epithelial tissue. (Kyle J Napier et al.,2014)

Columnar-lined esophagus (CLE, additionally referred to as Barrett's esophagus) is a common medical condition afflicting the human esophagus and is characterised by the replacement of the squamous epithelium in the lower esophagus by columnar epithelium. It's been well established that CLE is the premalignant condition from which EAC arises almost exclusively. Since EAC has a very poor prognosis, it's of great importance to understand the pathologic process and to develop methods for the interference of this deadly disease.(Xiaoxin Chen et al., 2001)

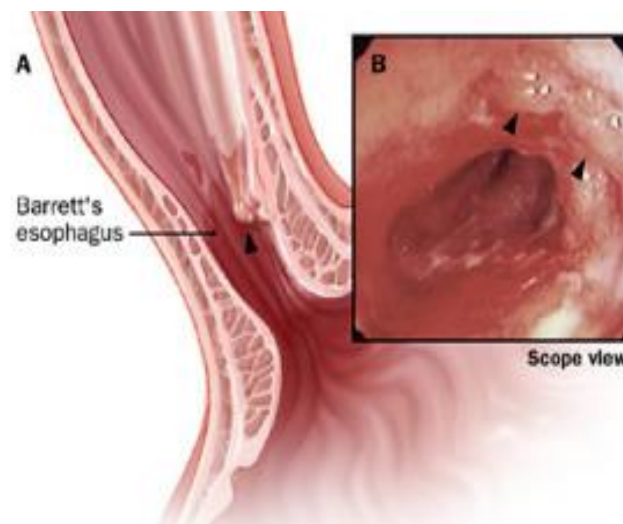


Fig1.2 Esophageal cancer

(https://www.hopkinsmedicine.org/gastroenterology_hepatology/_pdfs/esophagus_stomach/esophageal_cancer.pdf)

1.3.3 PANCREATIC

The pancreatic tumor microenvironment contains associate plentiful fibrotic stroma, which incorporates a spread of cell types and extracellular matrix (ECM)

components, like collagen, fibronectin, hyaluronic acid, and N-acetyl-glucosamine. The stroma is not simply a barrier for cancer cells and it is important in an exceedingly type of cellular processes, together with neoplasm formation, invasion, metastasis, and drug resistances

Stroma is related to the modification of cancer cell metabolism, immune cell accomplishment and also the regulation of acinar-to-ductal metaplasia within the progression.(Huiyun Zhu et al.,2018)

Pancreatic ductal adenocarcinoma (PDAC) is the one of the most lethal forms of cancer. In most of the cases PDAC involves activating KRAS mutations. In addition to driving growth, KRAS induces metabolic changes as well as increased glucose uptake, glycolytic flux, and glucose flux into hexosamines and ribose-5-phosphate. In contrast to alternative driver oncogenes like PI3K that broadly increase glucose flux throughout metabolism, oncogenic RAS impairs flux of glucose through pyruvate dehydrogenase into the TCA cycle. RAS-driven cells instead depend heavily on amino-alkanoic acid as a TCA carbon supply, with amino-alkanoic acid organic process through the TCA cycle and malic catalyst essential in carcinoma cells. Thus, RAS-driven cancer cells area unit relatively less captivated with aldohexose than alternative cancer cells(Jurre J. Kamphorst et al., 2015)

Pancreatic ductal adenocarcinoma (PDA) is fatal and therapy-resistant cancer that thrives in an exceedingly extremely desmoplastic, nutrient-deprived microenvironment. Many studies investigated the consequences of depriving PDA of either glucose or glutamine alone. However, the implications on PDA growth and metabolism of limiting both preferred nutrients have remained largely unknown. Mechanistically, adapted clones share common transcriptional and metabolic programs, with amino acid use for de novo glutamine and nucleotide synthesis.(Pei-Yun Tsai et al.,2020)

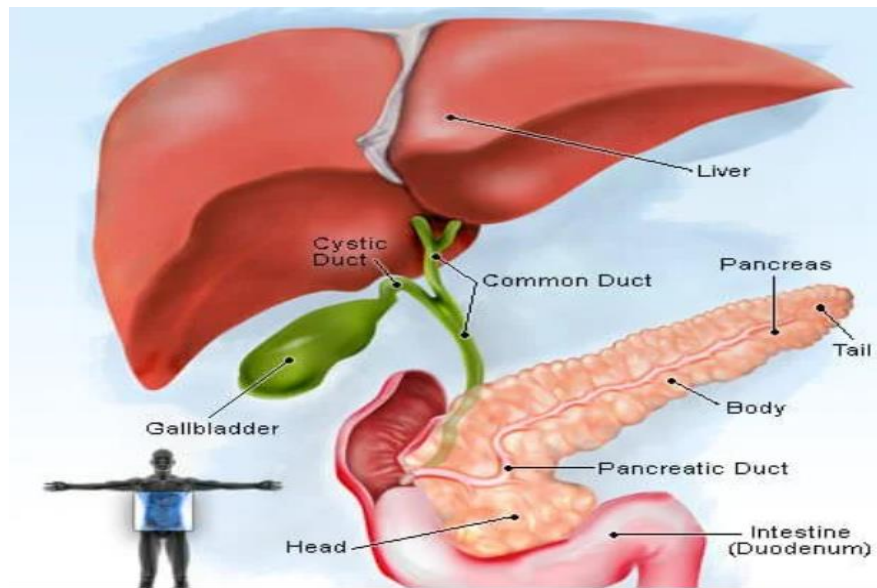


Fig1.5 Pancreatic cancer

(https://www.medicinenet.com/pancreatic_cancer/article.htm)

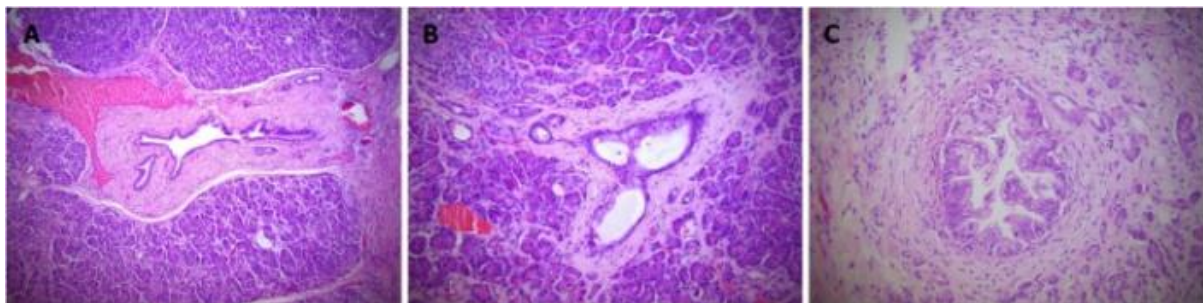


Fig 1.5.1 Stages of pancreatic cancer

(<https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6250924/>)

1.3.4 Prostate:

Prostate cancer ought to be suspected in men over fifty years old presenting with lower tract symptoms (LUTS), visible symptoms or dysfunction. LUTS are a typical presenting symptom of benign conditions touching the prostate, like benign prostate

dysplasia (BPH) and inflammation, making a diagnostic challenge.(Samuel W. D. Merriel et al.,2018)

Prostate cells have a definite metabolic profile reflective of the assembly of turn, prostate specific matter (PSA) and polyamines that are the major parts of prostate fluid. Once the prostate cells bear growth transformation they lose the capability to accumulate atomic number and end up in reconditioned activity of the mitochondrial catalyst m-aconitase and so turn reaction. Consequently turn levels decrease and nucleotide generation is enlarged. Furthermore a number of key enzymes concerned in carboxylic acid and steroid alcohol synthesis the measures of upregulated in adenocarcinoma cells(Sebastian Meller et al.,2016)

Primary prostate neoplasm heterogeneity is poor in initiation and advancement of the disease. The growth of neoplasm cells is accompanied by mutations in nuclear and in mitochondrial genomes. Thus, mitochondrial DNA mutations are also used as neoplasm cell markers. By using the laser capture microdissection which includes assays for mitochondrial genetic mutation detection, mtDNA mutations were used to trace mutated cells at a microscopic anatomy level.(Christian Arstad et al., 2020)

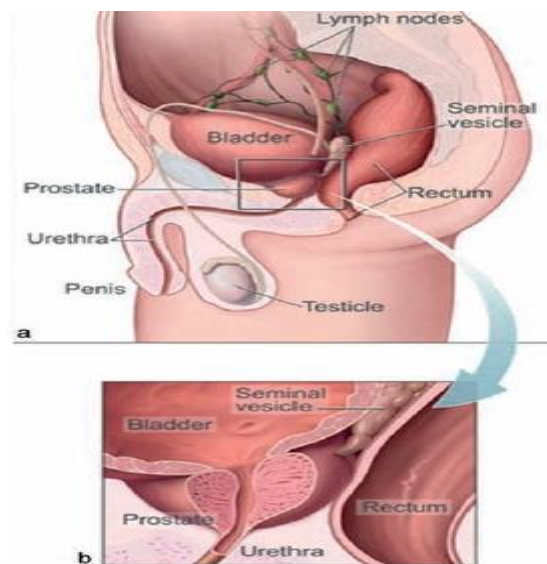


Fig1.6 Prostate cancer

(<https://www.cancer.gov/cancertopics/wyntk/prostate/allpages#ab3d4f20-6ab9-4428-9717-067035d2e691.htm>.)

1.3.5 Colorectal:

Colorectal cancer is the most common cancer type. The rise of colorectal cancer in developed countries are often attributed to the more and more ageing population, unfavourable trendy dietary habits and a rise in risk factors like smoking, low work up and fatness.(Ernst J. Kuipers et al.,2015)

Most colon tumors develop a series of histologic, morphological, and genetic changes that accumulate over time. This is done by screening and detection of early-stage malignant tumor polyps before they become cancerous in people at average risk for large intestine cancer (CRC), which can result in substantial decreases within the incidence of CRC.(Karen Simon.,2016)

CRC is caused by mutations that focus on oncogenes, tumor suppressor genes and genes associated with deoxyribonucleic acid repair mechanisms. Depending on the origin of the mutation, large intestine carcinomas is classified as unpredictable, inheritable and familial . The pathogenic mechanisms of this situation are enclosed in 3 sorts, namely chromosomal instability (CIN), microsatellite instability (MSI) and CpG island methylator phenotype (CIMP).(Inés Mármol et al.,2017)



Fig 1.7 Colorectal cancer at different stages

<https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4874655>)

1.3.6 Gastric cancer:

Gastric cancer is an aggressive disease that continues to have a frightening impact on international health. Incidence is declining because of improved nutrition, food preservation, better prevention, earlier diagnosing and treatment, the disease still carries a poor prognosis. stomachal cancer is commonly diagnosed at a sophisticated stage.(Lauren Peirce Carcas., 2014)

Gastric cancer (GC) is a complex disease, wherever several factors will influence its development, both environmental and genetic. Cancer incidents may occur by environmental agents, mostly dietary habits and social behavior. The event and progression of tumors may be a multiannual and time period method. Cancer usually happens by exposure to damaging malignant neoplastic disease agents.(Julita Machlowska et al.,2020)

Gastric cancer results from a mixture of environmental factors and accumulation of specific genetic alterations. The primary hindrance includes a healthy diet, anti-H. pylori therapies, chemoprevention, and screening for early detection. Dietary factors have a crucial impact on gastric carcinogenesis, particularly just in case of intestinal glandular cancer. Healthy dietary habits, that is, high intake of fresh fruits and vegetables, Mediterranean diet, a salt-free diet, salt-preserved food, red and high cured meat, sensible alcohol drinking, and maintaining a proper weight could be related to a minimized risk of GC.(Robert Sitarz et al.,2018)

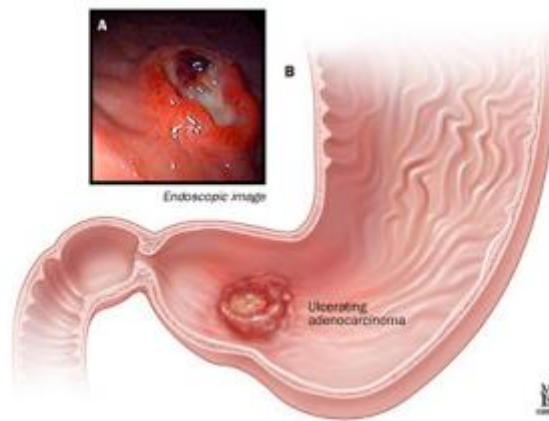


Fig 1.8 Gastric cancer

(https://www.hopkinsmedicine.org/gastroenterology_hepatology/_pdfs/esophagus_stomach/gastric_cancer.pdf)

11.4 INCIDENCE WORLD WIDE:

Cancer incidence and mortality are quickly growing worldwide. The reasons are advanced however reflect both aging and growth of the population, similarly as changes within the prevalence and distribution of the most risk factors for cancer, many of those are related to socioeconomic development. With fast increase and aging worldwide, the rising prominence of cancer as a leading reason for death partially reflects marked declines in mortality rates of stroke and coronary heart condition, relative to cancer, in several countries. (Freddie Bray.,2018)

Cancer's rising prominence as a number one reason for death part reflects marked declines in mortality rates of stroke and coronary heart condition, relative to cancer, in several countries. Overall, the burden of cancer incidence and mortality is rapidly growing worldwide; this reflects both aging and growth of the population yet as changes within the prevalence and distribution of the most risk factors for cancer, many of that are related to socioeconomic development. (Hyuna Sung et al.,2021)

Cancer could be a major public pathological state worldwide and is that the second leading explanation for death. A comprehensive overview of cancer prevalence

supported the foremost method for cancer incidence through 2015 and for mortality through 2016. we additionally estimate the whole range of deaths averted owing to the continual decline in cancer death rates since the early 1990s and analyze cancer mortality rates by county-level economic condition.(Rebecca L. Siegel et al.,2019)

1.5 GENE EXPRESSION OMNIBUS:

The GEO database (<http://www.ncbi.nlm.nih.gov/geo/>) is a global public repository that archives and freely distributes high-throughput gene expression and alternative functional genomics data sets.

Created in 2000 as a worldwide resource for gene expression studies, GEO has evolved with quickly dynamic technologies and currently accepts high-throughput knowledge for several alternative data applications.(Emily Clough et al., 2016)

GEO provides a versatile and open style that facilitates submission, storage and retrieval of heterogeneous data sets from high-throughput gene expression and genomic cross experiments. GEO isn't meant to switch in house gene expression intended bases of coherent data sets, and that are created to facilitate a particular analytic technique, however rather complement these by acting as a tertiary, central data distribution hub. The 3 central data entities of GEO are platforms, samples and series, and were designed with gene expression and genomic cross experiments.(Ron Edgar et al., 2002)

The flexible architecture of GEO results in a versatile design capable of accommodating heterogeneous data. GEO is not cancer-specific and attempts to encompass profiles from all studies using microarray technologies. In addition to data obtained from microarray methodologies, GEO also contains data types obtained from non-array-based techniques, such as serial analysis of gene expression (SAGE) (Barrett T, Edgar R, 2006).

1.6 DIFFERENTIAL EXPRESSED ANALYSIS

Differentially Expressed (DE) genes are defined as the genes that show differences in expression level between conditions, that is, the fraction of DE genes that were up-regulated and down-regulated, respectively, in one condition compared to the other. The identification of genes that are expressed significantly in different quantities in distinct groups of samples, be it biological condition diseased vs. healthy individuals, different tissues, and different stages of development. DEG involves the measurement of genes whose expression differ according to phenotype or experimental conditions. Some DEG analysis will not only measure variation but also the dependencies between genes. This requires the statistical analysis of large numbers of genes, to discover quantitative changes in expression levels between experimental groups. Microarrays have been used routinely for differential expression analysis for over a decade (<https://www.nih.gov/>).

The analysis of single-cell RNA sequencing (scRNAseq) information plays a very important role in understanding the intrinsic and external cellular processes in biological and medicine analysis. One vital effort during this space is that the detection of differentially expressed (DE) genes. The new approaches beyond the conventional ones, that are supported a nonzero distinction in average expression. Many strategies are developed for differential organic phenomenon analysis of scRNAseq information. (Tianyu Wang et al.-2019)

Ion channels and transporters (ICT) are increasingly rising as a novel category of membrane proteins expressed in many sorts of human cancers and regulation completely different aspects of neoplastic cell behavior. differing kinds of ICT are found to be functionally expressed in different sorts of cancer cells, and to manage completely different aspects of neoplasm cell behaviour. Seven cancer varieties (breast, prostate, lung, esophagus, stomach, colon and pancreas) that really represent great health issues, because of either high incidence or mortality rates.(ElenaLastraioli et al.,2015)

Significant analysis for screening DEG's was performed based on Student's t-test and Fold Change (FC) were used to identify the DEG's between normal and diseased samples.

P-value

The p-value is defined as an estimated probability of rejecting null hypothesis (H_0), when that hypothesis is true. In our context, null hypothesis was no differences between the normal and diseased samples for a particular gene i.e. If the p-value is less than significance level (<0.05), there we rejected the null hypothesis i.e. accepted a difference exists between normal and diseased sample for a particular gene (Ashwini Priya Jeggari, 2013).

Multiple Testing Problem (False Discovery Rate)

When the p-value threshold was set to 0.05, it means there is a 5% chance that the result is a false positive. While 5% was acceptable for one test; if we had done multiple tests, then this 5% would result in a large number of false positives which is called multiple testing problems. BenjaminiHochberg method (Benjamini Y, Hochberg Y, 1995) is to be applied to control FDR and accepted the cut-off 0.05 of this adjusted p-value (q-value), which implies that 5% of DE cases might be false positives in reality (Ashwini Priya Jeggari, 2013).

Fold change

Fold change estimates represent the biological significance and are often used in gene expression analysis of microarray experiments to measure the change in mRNA/gene expression level. It is a number in logarithm based 2 showing how many folds the gene expression has gained from normal(x) to diseased condition(y) (Ashwini Priya Jeggari, 2013).

Fig1.1 Statistical significance criteria

Statistical significance criteria	upregulated	downregulated
P-value	<0.05	>0.05
FC-value	>2	<2

Table 1.2 Parameters used for DEG screening

Parameter	Definition of the parameter	Reference
adj.P-Val	P-value after adjustment for multiple testing.	Wright, S. P, 1992
P-Value	Raw P-value	G. K. Smyth, 2005
P	Moderated t-statistic	G. K. Smyth, 2003
Log FC	Log2-fold change between two experimental conditions	Ashwini Priya Jeggari, 2013
F	Moderated F-statistic combines the t-statistics for all the pairwise comparisons into an overall test of significance for that gene	G. K. Smyth, 2005

1.7 HOW DEG ANALYSIS DONE

Differential expression analysis means taking the normalised read count data and performing statistical analysis to discover quantitative changes in expression levels between experimental groups. GEO2R is an interactive web tool that allows users to compare two or more groups of Samples in a GEO Series in order to identify genes that are differentially expressed across experimental conditions. The data obtained from GEO accession viewer was analysed by using GEO2R tool .Using the GEO2R defined the obtained GSM files into groups which may be 2 or multiple groups (<https://www.ncbi.nlm.nih.gov/geo/info/geo2r.html>).

1.8 ABOUT CYTOSCAPE

Cytoscape is an open source software project for integrating biomolecular interaction networks with high-throughput expression data. Cytoscape is most powerful when utilized in conjunction with large databases of protein-protein, protein-DNA, and genetic interactions that are increasingly available for humans and model organisms.(Paul Shannon et al., 2003) Cytoscape creates networks, wherever nodes of the network represent objects(such as proteins) and connecting edges represent relationships between them (such as physical interactions).Each Edge connects two Nodes. Edges will be directed or undirected. in the case of a directed edge, there is a Source and a Target Node.Once this basic network is created, numerous attributes of the nodes and edges will be additional to the network and incorporated as visual cues like form or color.(John Scooter Morris et al.,2012)

1.9 CYTOSCAPE USAGE IN NETWORK ANALYSIS

Cytoscape is widely used in biological network analysis and it supports many cases in molecular and systems biology, genomics and proteomics for biological networks representation and analysis of these interaction graphs allows us to map and

characterize the signalling pathways and predict the function of unknown proteins (Porras, Pablo, 2017). We can measure nodes by their network features to infer their

importance in the network, and it can help us identify central elements of biological networks. Network analysis can help us understand the function of an individual node and the collaboration between other nodes.

1.10 HUB GENE ASSESSMENT

Due to the development of high-throughput microarray technology, associate increasing variety of genes are known to serve a vital role in tumour prevalence and within the progression of LUAD. gene expression profiles were used to determine vital genes related to tumour progression. However, the bulk of studies have targeted on differentially expressed genes (DEGs) and not on the interconnection between genes. Therefore, tumour stage was selected as a main clinical feature. Later, co-expression networks of the association between genes were made, and network-centric genes related to the clinical options were known.(Dong-Hu Yu et al., 2019)

The high mortality of colorectal cancer (CRC) patients and also the limitations of standard tumor-node-metastasis (TNM) stage emphasized the need of exploring hub genes closely associated with carcinogenesis and prognosis in CRC. The DEGs were any preformed functional enrichment analysis to analyze biological processes, molecular functions, and reactome pathways regulated by the DEGs. The protein-protein interaction (PPI) network reflective the interactions among DEGs was made, and hub network modules were captured and deciphered, that embodied representative genes in CRC carcinogenesis.(Benjiao Gong et al.,2020)

The specific symptoms and pathological images may only be detected unambiguously at the late stage of carcinoma. This raises an imperative need for the development of reliable biomarkers which may effectively differ tumor from normal tissues supported analyses of gene expression profiles. In order to spot novel

diagnostic predictors and molecular markers, we tend to integrate 2 microarray datasets from GEO info, and 11 candidate genes considerably differentially

expressed between tumor and normal samples were screened by bioinformatics analyses.(Chunyang Li et al.,2018)

1.2 REVIEW OF LITERATURE:

On 2019, Cheng Yue et al has described that lung cancer has the highest morbidity and mortality worldwide. lung adenocarcinoma (LADC) is the commonest pathological subtype. Accumulating proof suggests that the tumour microenvironment (TME) is related with the tumour progress and therefore the patient's outcome because the major elements of TME, the tumor-infiltrated immune cells and stromal cells have attracted more attention. This study aims at differentially expressed immune and stromal signature genes that are employed to construct a TME-related prognostic model for predicting the outcomes of LADC patients.

On 2020, Chao Ma et al has described that lung adenocarcinoma (LUAD) is the most diagnosed histological subtype of NSCLC, followed by squamous cell cancer. Understanding the molecular pathological process of LUAD that provides necessary data concerning prognosis. The surroundings within which the growth exists is termed the growth microenvironment and consists of fluids, extracellular matrix molecules, immune cells, stromal cells, and inflammatory cells. The growth microenvironment is extremely dynamic and it is important in jointly promoting the immune escape, growth, and metastasis of tumors. Levels of immune cell infiltration are according to be associated with prognosis, and also the activity of each immune and stromal cells has been shown to predict overall cancer survival. Each component of the microenvironment plays a crucial role in LUAD, and so contains a vital impact on clinical outcomes.

On 2020, Yan Qu et al has described non-small cell lung cancer (NSCLC), which consists primarily of lung adenocarcinoma (LUAD) and lung squamous cell carcinoma (LUSC), that results in cancer deaths worldwide.

This study performs the comprehensive analysis of growth microenvironmental and genetic factors to spot prognostic biomarkers for NSCLC. supported these scores, the differentially expressed genes were obtained and immune-related prognostic genes were known. functional analysis and protein-protein interaction network further revealed the immune-related biological processes during which these genes participated. many differentially expressed genes were known to be potential immune-related prognostic genes. This study aims to comprehensively analyze the growth microenvironment and conferred immune-related prognostic biomarkers for NSCLC.

On 2020, Chao Ma et al described the aim to establish a gene signature based on tumour microenvironment which will predict patients' outcomes for LUAD. An eight-gene signature was designed, and it was examined by Kaplan–Meier analysis, revealing the numerous overall survival differences. GSEA and immune infiltrating analyses showed that the precise pathways associated with the characteristics of eight-genes signature, and known the important roles of Mast cells resting and B cells naive within the prognosis of the eight-gene signature.

On 2021 Juan Chen et al described the lung adenocarcinoma (LUAD) as the most common histological form of lung cancers, that is the primary reason behind cancer-related mortality worldwide. Growing proof has recommended that tumour microenvironment (TME) plays a crucial role in tumorigenesis and progression. Hence, we tend to investigate the correlation of TME connected genes with LUAD prognosis.

CHAPTER-3

3.1 AIM:

Identification of differentially expressed genes and key genes associated with adenocarcinoma

3.2 OBJECTIVE:

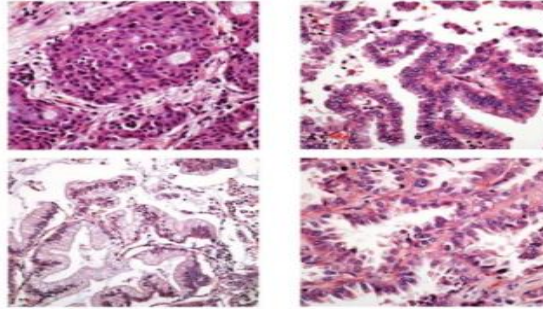
- Literature survey on gene expression profile of adenocarcinoma in lung
- Identification of Differentially Expressed Genes (DEG's) in lung cancer, esophageal cancer, prostate cancer, colorectal cancer, pancreatic cancer, gastric cancer.
- Hub Gene assessment based on degree in order to evaluate the importance of nodes in biological networks.
- Prognosis analysis of Differentially expressed genes is used to analyse the gene signature.
- Identification of DEG associated with adenocarcinoma

3.3 SCOPE OF THE WORK:

The identified DEG's and hub genes may guide investigations on the mechanism underlying the development and progression of lung cancer, lung cancer, esophageal cancer, prostate cancer, colorectal cancer, pancreatic cancer, gastric cancer associated with adenocarcinoma issues. Further experimental studies are necessary to confirm the present findings.

CHAPTER-4

MATERIAL AND METHODS



Literature survey on adenocarcinoma in lung, esophageal, prostate, pancreatic, colorectal, gastric cancer occurrence



Samples		Define groups		Selected 10 out of 10 samples						
Group	Accession	Enter a group name:	List	Sample name	Tissue	Status after 5 years	Histology	Age	Gender	Stage
primary lung ...	GSM4040012	Cancel selection		cancer tissue specimen	primary lung adenocarcinoma	Dead	AD	65 years	Female	IA
primary lung ...	GSM4040013	primary lung adenocarcinoma (5 samples)		cancer tissue specimen	primary lung adenocarcinoma	Alive	AD	60 years	Female	IIB
primary lung ...	GSM4040014	non-tumor lung tissue (5 samples)		cancer tissue specimen	primary lung adenocarcinoma	Alive	AD	54 years	Male	IIB
primary lung ...	GSM4040015			cancer tissue specimen	primary lung adenocarcinoma	Dead	AD	57 years	Male	IIIA
primary lung ...	GSM4040016	0030340590T		fresh lung cancer tissue specimen	primary lung adenocarcinoma	Alive	AD	52 years	Female	IB
non-tumor lung ...	GSM4040007	568369N		non-tumor lung tissue specimen	non-tumor lung tissue	Dead	AD	57 years	Male	IIIA
non-tumor lung ...	GSM4040008	0001156551N		non-tumor lung tissue specimen	non-tumor lung tissue	Dead	AD	65 years	Female	IA
non-tumor lung ...	GSM4040009	0001162111N		non-tumor lung tissue specimen	non-tumor lung tissue	Alive	AD	60 years	Female	IIB
non-tumor lung ...	GSM4040010	0001164648N		non-tumor lung tissue specimen	non-tumor lung tissue	Alive	AD	52 years	Female	IB
non-tumor lung ...	GSM4040011	0030330793N		non-tumor lung tissue specimen	non-tumor lung tissue	Alive	AD	54 years	Male	IIB

Gene expression profile retrieval - Gene Expression Omnibus(GEO)



Differential expression analysis of cancer datasets related to adenocarcinoma condition of lungs, gastric, esophageal, prostate, pancreas,colorectal.



Screening of upregulated and down-regulated genes pvalue (<0.05) and FC value (>2) were listed as upregulated and FC <2 and P value >0.05 were listed as downregulated genes



Hub gene assessment was done using Cytoscape plugin - Cytohubba



Screening of marine based compounds binding affinity towards the hub gene GAL3



Marine based compounds docking and intermolecular interaction assessment

Fig 4.1 Proposed work plan

4.1 LUNG CANCER, ESOPHAGEAL CANCER, PROSTATE CANCER, COLORECTAL CANCER, PANCREATIC CANCER, GASTRIC CANCER EXPRESSION DATASETS:

To explore the differentially expressed genes between lung cancer, esophageal cancer, prostate cancer, colorectal cancer, pancreatic cancer, gastric cancer raw gene expression profile data were downloaded from Gene Expression Omnibus Database(GEO,<https://www.ncbi.nlm.nih.gov/geo>), is a repository of public functional genomics data. GEO accession numbers: GSE136043 (Lung), GSE92396 (Esophageal), GSE 69223 (prostate), GSE62452 (pancreatic), GSE113513 (colorectal), and GSE103236 (gastric). GSE136043 dataset was submitted by Jiang N et al., 2019, GSE92396 dataset was submitted by El-Rifai W et al., 2016, GSE 69223 dataset was submitted by Zhen-yu Song et al., 2015, GSE62452 dataset was submitted by Yang S et al., 2014, GSE113513 dataset was submitted by Jun Peng 2018, and GSE103236 dataset was submitted by Chivu Economescu M et al., 2017 respectively. The gene expression profile analysis of GSE136043 and GSE113513 was performed based on GPL15207 platform (Prime View – Affymetrix Human Gene Expression Array) and sample size is 28 and 28 respectively. The sample size of GSE92396 and GSE62452 is 22 and 130 respectively which is performed based on GPL6244 platform([HuGene-1_0-st] Affymetrix Human Gene 1.0 ST Array[transcript (gene) version]), GSE 69223 was performed based on GPL570 platform (microarray) which contains 15 sample size, and the sample size of GSE103236 is 19 which performed based on GPL41333 platform (Agilent-014850 Whole Human Genome Microarray 4x44KN G4112F (Feature Number version))

Fig4.2 GEO dataset associated with adenocarcinoma DEG analysis

Dataset ID	Sample	GPL ID
GSE136043	10	GPL15207(Prime View – Affymetrix Human Gene Expression Array)
GSE113513	28	GPL15207(Prime View – Affymetrix Human Gene Expression Array)
GSE92396	22	GPL6244 ([HuGene-1_0-st] Affymetrix Human Gene 1.0 ST Array[transcript (gene) version])
GSE62452	130	GPL6244 ([HuGene-1_0-st] Affymetrix Human Gene 1.0 ST Array[transcript (gene) version])
GSE 69223	15	GPL570 (Affymetrix Human Genome U133 Plus 2.0 Array)
GSE103236	19	GPL41333 (Agilent-014850 Whole Human Genome Microarray 4x44KN G4112F (Feature Number version) .

4.2 IDENTIFICATION OF DIFFERENTIALLY EXPRESSED GENES (DEG's):

The identification of differentially expressed genes (DEG's) between lung cancer and normal lung tissues was performed using GEO2R(<http://www.ncbi.nlm.nih.gov/geo/geo2r/>), a web tool, can perform sophisticated R-based analyses of GEO data and presents the results as a table of differential gene expression that can be visualized using GEO Profile graphics. Significant analysis was performed based on Student's t-test and Fold Change (FC) were used to identify the DEG's between lung cancer and normal lung tissue. Then in order to calculate the adjusted P-values (False Discovery Rate [FDR]) Benjamini and Hochberg method was used to minimize errors due to multiple hypothesis testing. Among these differentially expressed genes which includes both up-regulated and down-regulated, only the genes with $\log_{2}FC > 2$ and adjusted P-value < 0.05 were set as the criterion of statistical significance for identifying the significant upregulated DEG's. By comparing the list of DEG's, common genes between lung cancer, esophageal cancer, prostate cancer, colorectal cancer, pancreatic cancer, gastric cancer.

4.3 PROTEIN-PROTEIN INTERACTION NETWORK CONSTRUCTION

Protein–protein interactions (PPIs) are very necessary in orchestrating the events during a cell. They form the basis for many signal transduction pathways during a cell, yet as numerous transcriptional regulatory networks(Karthik Raman., 2010) PPI network analysis is an attractive technique which is used for molecular assessment of many diseases (Shannon et al., 2003; Massoud and Gambhir, 2007). In this method, many proteins implicated to a certain disease are retrieved and attributed in an interactive structure (PPI network) (Liu et al., 2014; Safari-Alighiarloo et al., 2014; Abbaszadeh et al., 2017). Topological analysis of networks provide useful information about roles of the initial nodes in the network (Safari-Alighiarloo et al., 2014; Moreau and Tranchevent, 2012; Safari-Alighiarloo et al., 2016). Crucial nodes of the network are the important proteins related to the studied disease (Safari-Alighiarloo et al., 2016; Zamanian-Azodi et al., 2015). The Proteomics

Standard Initiative Common QUery InterfaCe (PSICQUIC) enables computational access to molecular-interaction data resources by means of a standard Web Service and query language. PSICQUIC interface allows the concurrent search of multiple molecular-interaction information resources from 28 servers globally (BioGRID, REACTOME, DIP, HPRD, IntAct, MINT, SPIKE, iRefIndex etc.,) which provides greater than 150 million binary interactions using a single query in MIQL (Molecular Interactions Query Language) (Del-Toro, N., *et al.*, 2013).

BioGRID database was used for construction of an integrated PPI network for the DEG's. The Biological General Repository for Interaction Datasets (BioGRID) is an open access database dedicated to the annotation and archival of protein, genetic and chemical interactions for all major model organism species and humans. BioGRID curates the biomedical literature for major model organism species, including humans, with a recent emphasis on central biological processes and specific human diseases. A new dynamic interaction network viewer allows the easy navigation and filtering of all genetic and protein interaction data, as well as for bioactive compounds and their established targets (Andrew Chatr-aryamontri *et al.*, 2017). As the PPI network reflects the mode of action caused by interruptions of the protein and related targets. Crucial PPIs are proven to participate in disease-associated signalling pathways, which can offer insight for novel target identification and drug discovery.

4.4 NETWORK ASSESSMENT TO SCREEN HUBS:

Nodes with high degree, known as hubs, can be key players in molecular mechanisms such as a protein interacting with multiple other proteins, regulation of multiple genes by a key transcription factor or multipart regulation by other regulatory elements [i.e. microRNAs (miRNAs)]. In human cells, hub genes have been found to indicate essential genes (i.e. critical for survival) rather than disease genes. Since the biological network is

heterogeneous, it is reasonable to use more than one method for catching essential proteins. Thus, these essential nodes/hubs may serve as potential drug-targets for developing novel therapy of human diseases, such as cancer (Chia-Hao Chin, Shu-Hwa Chen et al., 2008).

A novel Cytoscape plugin CytoHubba which provides a user-friendly interface to explore important nodes in biological networks is used to perform hub gene assessment by ranking nodes in a network by their network features. Thus cytoHubba application helps to search and explore the network and to extract user interesting subnetwork.

CytoHubba provides 11 topological analysis methods. These 11 methods can be divided into two major categories: Local and Global Based Methods. To calculate the score of a node within a network, a local rank method only considers the relationship between the node and its direct neighbors; on the other hand, the global method examines the relationship between the node and the entire network. There are four Local based methods namely Degree method, Maximum Neighborhood Component (MNC), Density of Maximum Neighborhood Component (DMNC), Maximal Clique Centrality (MCC). Global based methods includes six centralities (Bottleneck, Eccentricity, Closeness, Radiality, Betweenness, and Stress) based on shortest paths and Edge Percolated Component (EPC). CytoHubba control panel is also a handy tool to retrieve sub networks from the whole big PPI network and fetches the shortest path among a group of nodes (Chia-Hao Chin et al., 2014).

4.5 MOLECULAR DOCKING

Docking is a method which predicts the preferred orientation of one molecule to a second when bound to each other to form a stable complex (Lengauer T, Rarey M 1996). In general, "docking" is the identification of the low-energy binding modes of a small molecule, or ligand, within the active site of a macromolecule, or receptor, whose structure is known. A compound that interacts strongly with, or binds, a receptor associated with a disease may inhibit its function and thus act as a drug. Solving the docking problem computationally requires an accurate representation of the molecular energetic as well as an efficient algorithm to search the potential binding modes. Two approaches are particularly popular within the molecular docking community. One

approach uses a matching technique that describes the protein and the ligand as complementary surfaces. The second approach simulates the actual docking process in which the ligand-protein pairwise interaction energies are calculated (Böhm HJ, 1998; Seifert MH *et al.*, 2007; Joseph-McCarthy D, 2007).

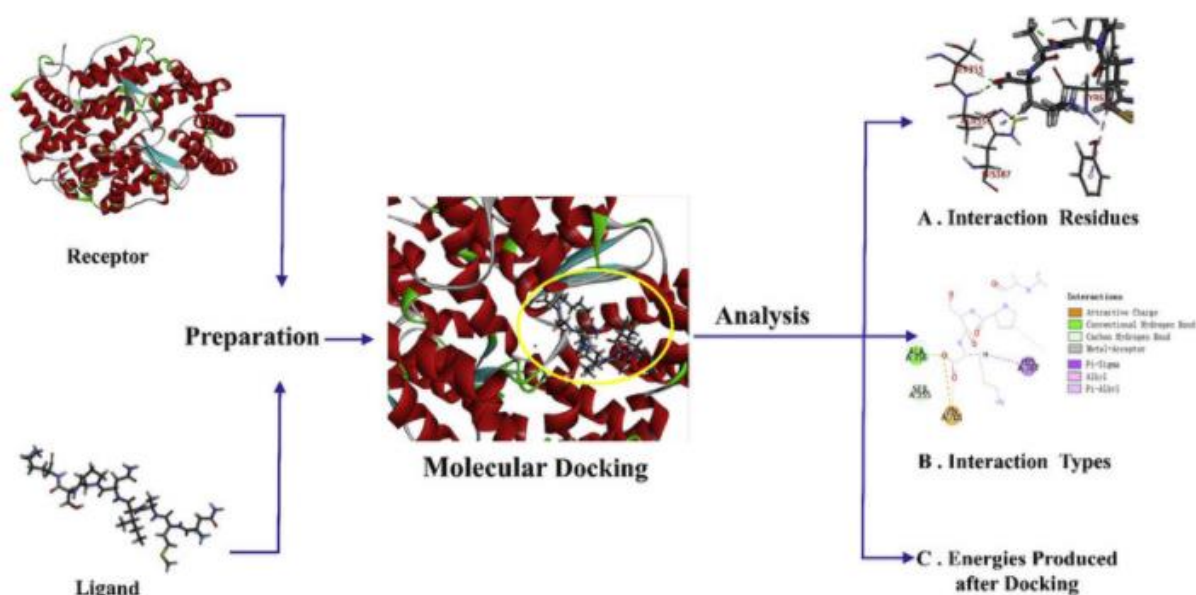


Fig4.2 Steps involved in molecular docking

(https://www.researchgate.net/figure/General-procedures-for-molecular-docking_fig3_324508414)

4.5.2 LigandFit docking

LigandFit was originally developed by Venkatachalam *et al.*, now it is implemented in the Discovery Studio software (version 2.0, Accelrys Software) as a molecular docking module. LigandFit docking algorithm is based on a shape comparison filter combined with a Monte Carlo conformational search method to

generate the docked poses which are consistent with the shape of the binding site followed by further CHARMM minimization. It randomly generates ten poses of the ligand in different orientations in the binding pocket. It then optimizes the poses by rigid body minimization of the ligand with respect to a calculated grid-based interaction energy using the Dreiding force field. The ten different poses are then ranked based on different scoring functions. The Dock score, combination of interaction energy of ligand and receptor together with the internal energy of the ligand, is the default scoring function for LigandFit docking program.

CHAPTER-5

RESULTS AND DISCUSSION

In the present study, Common DEG's among lung cancer, esophageal cancer, pancreatic cancer, prostate cancer, colorectal cancer, gastric cancer datasets were mined to reveal the common genes. PPI network was constructed for the overlapping DEG's and Hub genes were assessed based on the degree measures (eleven topological methods) available in CytoHubba plugin for screening the promising hub genes for genes associated with adenocarcinoma. Screening of Natural compounds from marine compounds antagonist effects will be analysed for their affinity towards the proposed therapeutic targets. Subsequently, drug likeness studies were carried out on the retrieved hits to evaluate and analyse their drug like properties such as absorption, distribution, metabolism, excretion, and toxicity (ADMET) for toxicity prediction. Then Molecular docking was carried out for natural, synthetic marine compounds.

5.1 IDENTIFICATION OF DEG's

With $\log_{2}FC > 2$ and adjusted P-value < 0.05 were set as the criterion of statistical significance for identifying the significant upregulated DEG's were shown in Table 5.2. We identified 622 Upregulated and 1536 downregulated genes in GSE136043 (LAUD), 465 upregulated and 5695 downregulated genes in GSE92396 (Esophageal Cancer), 45 upregulated and 15,891 downregulated genes in GSE62452 (Pancreatic Cancer), 254 upregulated and 6113 downregulated genes in GSE69223 (Prostate cancer), 787 upregulated and 15,376 downregulated genes in GSE113513 (Colorectal Cancer), 134 upregulated and 713 downregulated genes in GSE1032 (Gastric cancer) were shown in table 5.3.

Table 5.1 Datasets associated with adenocarcinoma DEG analysis

Cancer type	GEO accession number	Sample size	Normal	Diseased	Platform
Lung cancer	GSE136043	10	5	5	GPL15207(Prime View – Affymetrix Human Gene Expression Array)
Esophageal cancer	GSE92396	22	12	10	GPL6244 ([HuGene-1_0-st] Affymetrix Human Gene 1.0 ST Array[transcript (gene) version])
Prostate cancer	GSE69223	30	15	15	GPL570 (Affymetrix Human Genome U133 Plus 2.0 Array)
Pancreatic cancer	GSE62452	130	61	69	GPL6244 ([HuGene-1_0-st] Affymetrix Human Gene 1.0 ST Array[transcript (gene) version])

Colorectal cancer	GSE113513	28	14	14	GPL15207(Prime View – Affymetrix Human Gene Expression Array)
Gastric cancer	GSE103236	19	9	10	GPL41333 (Agilent-014850 Whole Human Genome Microarray 4x44KN G4112F (Feature Number version))

Table 5.2 Criterion of Statistical Significance for Identifying the Significant Upregulated DEG's.

Lung cancer	Esophageal cancer	Prostate cancer	Pancreatic cancer	Colorectal cancer	Gastric cancer
GSE136043 2158(DEG'S) p value<0.05 FC>2 Sample size: 10	GSE92396 6160(DEG'S) p value<0.05 FC>2 Sample size:22	GSE69223 547(DEG'S) p value<0.05 FC>2 Sample size:30	GSE62452 15936(DEG'S) p value<0.05 FC>2 Sample size:130	GSE113513 16163(DEG'S) p value<0.05 FC>2 Sample size:28	GSE103236 871(DEG'S) p value<0.05 FC>2 Sample size:19

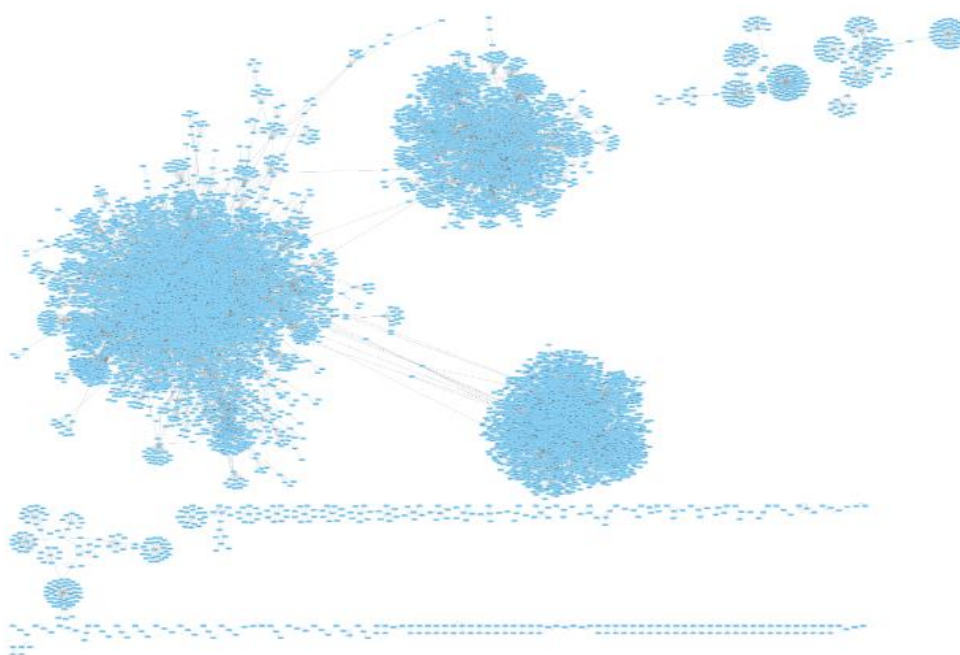
Fig 5.3 Upregulated and downregulated genes for the significant dataset

GEO Profile ID	Upregulated	Downregulated
GSE136043	622	1536
GSE92396	465	5695
GSE62452	45	15,891
GSE69223	254	6113
GSE113513	787	15,376
GSE103236	134	713

5.3 PPI NETWORK CONSTRUCTION AND ANALYSIS

Based on the predicted interactions of identified common DEG's, the PPI network was constructed by using BioGRID Database to identify the most important proteins that may serve crucial roles in the development of Adenocarcinoma. A total of 2688 genes (nodes) and 5340 gene-gene interactions (edges) respectively where each node represents a protein and the edge between the nodes represents the interaction between the proteins were mapped in the PPI network. Each gene was assigned a degree representing the number of neighboring nodes in the network and changes in the proteins/genes. A novel Cytoscape plugin *CytoHubba* which provides a user-friendly interface to explore important nodes in biological networks *is used to* perform hub gene assessment by ranking nodes in a network by their network features, helps to search and explore the network and to extract user interesting subnetwork. *CytoHubba* provides 11 topological analysis methods.

Fig 5.3 BioGRID Based Network With 2688 Nodes, 5340Edges



5.5 NETWORK ASSESSMENT TO SCREEN HUB GENES

The hub genes were selected from the PPI network using the 11 topological methods in CytoHubba plugin like Degree method. Each gene was assigned a degree representing the number of neighboring nodes in the network and changes in the proteins/genes. The top 10 hub nodes with the highest degrees in lung, prostate, esophageal, pancreatic, colorectal, gastric associated adenocarcinoma were GAL3, PARK11, UBC, PEPCK, IGHA1, RAI1, EXOSC7, APP, TCSF, PDB3, INSL5 were identified as the key genes. Among these hub genes GAL were chosen as the critical Hubs based on their rank among all the eleven topological methods. The high degree of these hub genes indicated that these proteins may serve crucial roles in maintaining the whole protein interaction network. Hence GAL3 were considered as promising biomarkers which could also serve as a therapeutic target for adenocarcinoma.

GAL3:

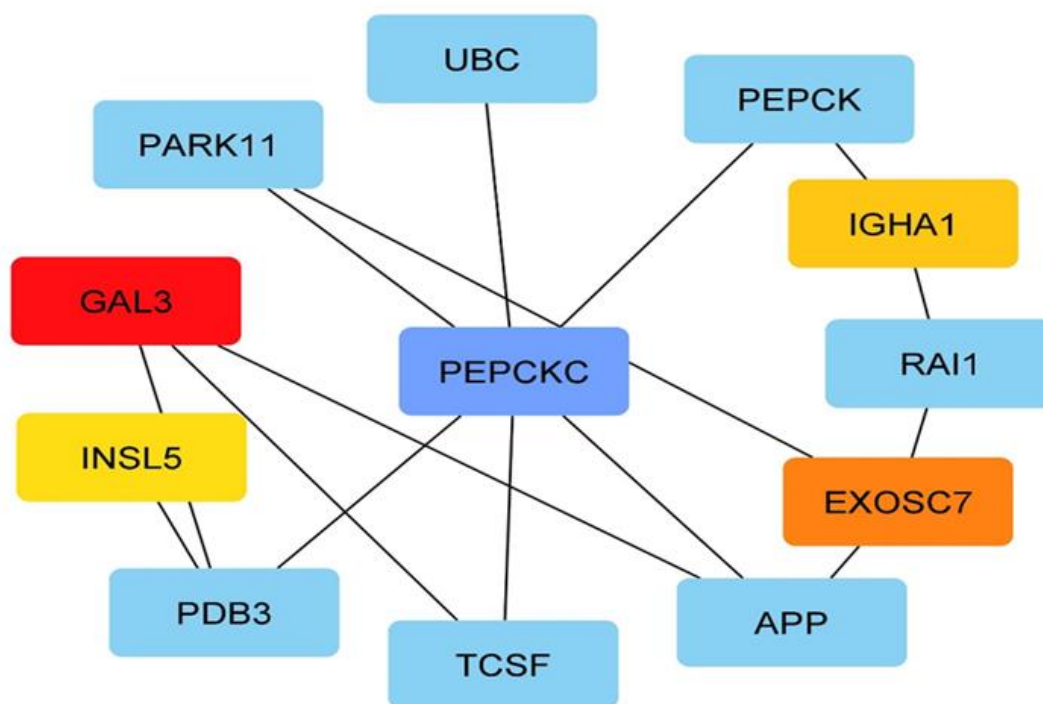
Galectins or S-type lectins is an evolutionarily preserved family of carbohydrate-binding proteins characterised by the presence of β -galactoside-binding sites, usually represented by 130 organic compound supermolecule recognition domain (CRD). These molecules are functionally concerned in many physiological processes, among that inflammation,

immune-response, RNA splice, sequence transcription, apoptosis, signaling, cell migration, and differentiation.(Armando Bartolazzi., 2018)

Galectins are soluble proteins synthesized on free ribosomes within the cytoplasm. They lack the classical characteristics of secreted proteins, they are rapidly translocated to the extracellular space through a yet unknown pathway. Once within the extracellular atmosphere, the galectins square measure exposed to a large type of glycan structures, wherever they recognize and bind specific β -galactosides.(John Stegmayr et al.,2019)

Galectin 3's ability to support neoplastic cell survival by intra-cellular and extra-cellular mechanisms suggests this molecule is a vital element of the tumour microenvironment that probably may be targeted for therapy. Galectin three is elevated in several cancers together with solid tumors and also the cancers of the blood. Galectin three conjointly seems to be a key molecule made by tumour microenvironment support cells together with mesenchymal stromal cells (MSC) to suppress immune surveillance by killing T cells and interfering with NK cell perform and by supporting metastasis.(Peter P.Ruvolo., 2016)

Fig 5.4 Hub Genes of GAL3



5.7 SCREENING OF MARINE BASED COMPOUNDS BINDING AFFINITY TOWARDS THE HUB GENE GAL3

Compounds from marine resources are taken in order to analyze their antagonist effects for their affinity towards the proposed therapeutic targets. Marine compounds considered for docking study and Galectin3 were screened for their affinity towards the proposed drug targets.

Table 5.4 - Docked Complex Of 6QLU with Compounds available Red Algae

Name	3D	2D
Red Algae		

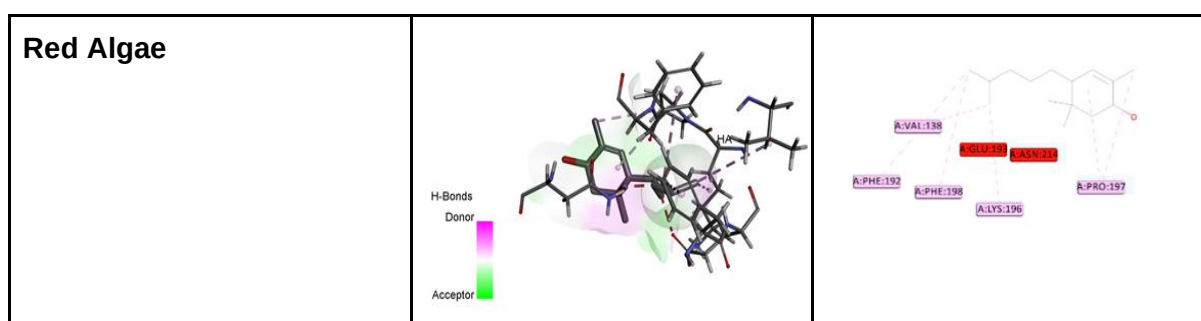


Table5.5- INTERACTION TABLE OF RED ALGAE

Name	Distance	Category	Types	From	From chemistry	To	To chemistry
A:HIS812 - Molecule:C3	4.94025	Hydrophobic	Pi-Alkyl	A:HIS812	Pi-Orbitals	Molecule:C3	Alkyl
Molecule:C3 - Molecule:C3	3.09214	Hydrophobic	Alkyl	Molecule:C3	Alkyl	Molecule:C3	Alkyl
Molecule:C3 - Molecule:C17	4.31786	Hydrophobic	Alkyl	Molecule:C3	Alkyl	Molecule:C17	Alkyl
Molecule:C7 - Molecule:C15	2.41574	Hydrophobic	Alkyl	Molecule:C7	Alkyl	Molecule:C15	Alkyl
Molecule:C7 - Molecule:C17	4.17457	Hydrophobic	Alkyl	Molecule:C7	Alkyl	Molecule:C17	Alkyl
Molecule:C8 - Molecule:C17	3.15874	Hydrophobic	Alkyl	Molecule:C8	Alkyl	Molecule:C17	Alkyl
Molecule:C13 - Molecule:C8	4.36028	Hydrophobic	Alkyl	Molecule:C13	Alkyl	Molecule:C8	Alkyl
Molecule:C14 - Molecule:C3	3.96327	Hydrophobic	Alkyl	Molecule:C14	Alkyl	Molecule:C3	Alkyl
Molecule:C14 - Molecule:C8	2.79174	Hydrophobic	Alkyl	Molecule:C14	Alkyl	Molecule:C8	Alkyl
Molecule:C14 - Molecule:C10	2.88083	Hydrophobic	Alkyl	Molecule:C14	Alkyl	Molecule:C10	Alkyl
Molecule:C14 - Molecule:C15	4.01119	Hydrophobic	Alkyl	Molecule:C14	Alkyl	Molecule:C15	Alkyl
Molecule:C14 - Molecule:C17	3.10655	Hydrophobic	Alkyl	Molecule:C14	Alkyl	Molecule:C17	Alkyl
Molecule:C15 - Molecule:C3	4.29314	Hydrophobic	Alkyl	Molecule:C15	Alkyl	Molecule:C3	Alkyl
Molecule:C15 - Molecule:C17	3.97115	Hydrophobic	Alkyl	Molecule:C15	Alkyl	Molecule:C17	Alkyl
Molecule:C3 - Molecule	4.72875	Hydrophobic	Alkyl	Molecule:C3	Alkyl	Molecule	Alkyl
Molecule:C8 - Molecule	4.99774	Hydrophobic	Alkyl	Molecule:C8	Alkyl	Molecule	Alkyl
Molecule:C10 - Molecule	3.6572	Hydrophobic	Alkyl	Molecule:C10	Alkyl	Molecule	Alkyl
A:HIS812 - Molecule:C3	4.94025	Hydrophobic	Pi-Alkyl	A:HIS812	Pi-Orbitals	Molecule:C3	Alkyl

Table5.6 - Docked complex of 6QLU with Green Algae

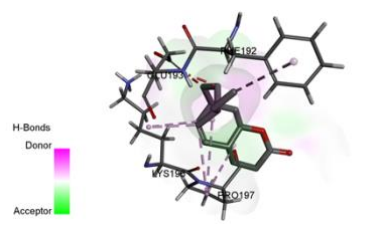
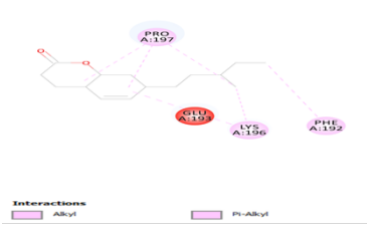
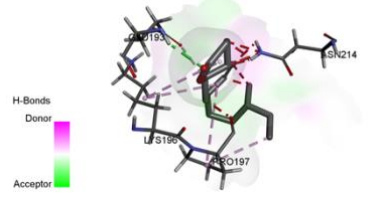
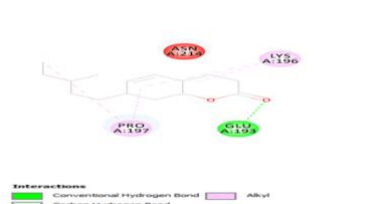
Name of the compound	3D	2D
Green Algae		
Green Algae		

TABLE 5.7 -INTERACTION TABLE OF GREEN ALGAE

Name	Distance	Bond	type	from	from chemistry	to	To Chemistry
A:PRO197 - Molecule	4.66858	Hydrophobic	Alkyl	A:PRO197	Alkyl	Molecule	Alkyl
Molecule:C7 - A:PRO197	4.68612	Hydrophobic	Alkyl	Molecule:	Alkyl	A:PRO197	Alkyl
Molecule:C14 - A:VAL138	5.11752	Hydrophobic	Alkyl	Molecule:	Alkyl	A:VAL138	Alkyl
Molecule:C14 - A:LYS196	3.49386	Hydrophobic	Alkyl	Molecule:	Alkyl	A:LYS196	Alkyl
A:PHE192 - Molecule:C15	5.11465	Hydrophobic	Pi-Alkyl	A:PHE192	Pi-Orbitals	Molecule:	Alkyl
A:PHE198 - Molecule:C15	5.17843	Hydrophobic	Pi-Alkyl	A:PHE198	Pi-Orbitals	Molecule:	Alkyl
A:LYS196:HZ2 - A:GLU193:OE2	2.0907	Hydrogen Bond;E	Salt Bridge;Attract	A:LYS196:I	H-Donor;Positive	A:GLU193:H	Acceptor;Ne
A:VAL138:HN - A:LYS196:O	2.18526	Hydrogen Bond	Conventional Hyd	A:VAL138:H	Donor	A:LYS196:H	Acceptor
A:LYS196:HN - A:GLU193:O	2.26132	Hydrogen Bond	Conventional Hyd	A:LYS196:I	H-Donor	A:GLU193:H	Acceptor
A:ASN214:HN - A:ASN214:OD1	2.81061	Hydrogen Bond	Conventional Hyd	A:ASN214	H-Donor	A:ASN214	H-Acceptor
A:LYS196:HE2 - A:GLU193:OE2	2.98096	Hydrogen Bond	Carbon Hydrogen	A:LYS196:I	H-Donor	A:GLU193:H	Acceptor
A:PHE198:HA - A:ASN214:OD1	2.5129	Hydrogen Bond	Carbon Hydrogen	A:PHE198	H-Donor	A:ASN214	H-Acceptor
A:PHE192 - A:PHE198	5.40777	Hydrophobic	Pi-Pi T-shaped	A:PHE192	Pi-Orbitals	A:PHE198	Pi-Orbitals

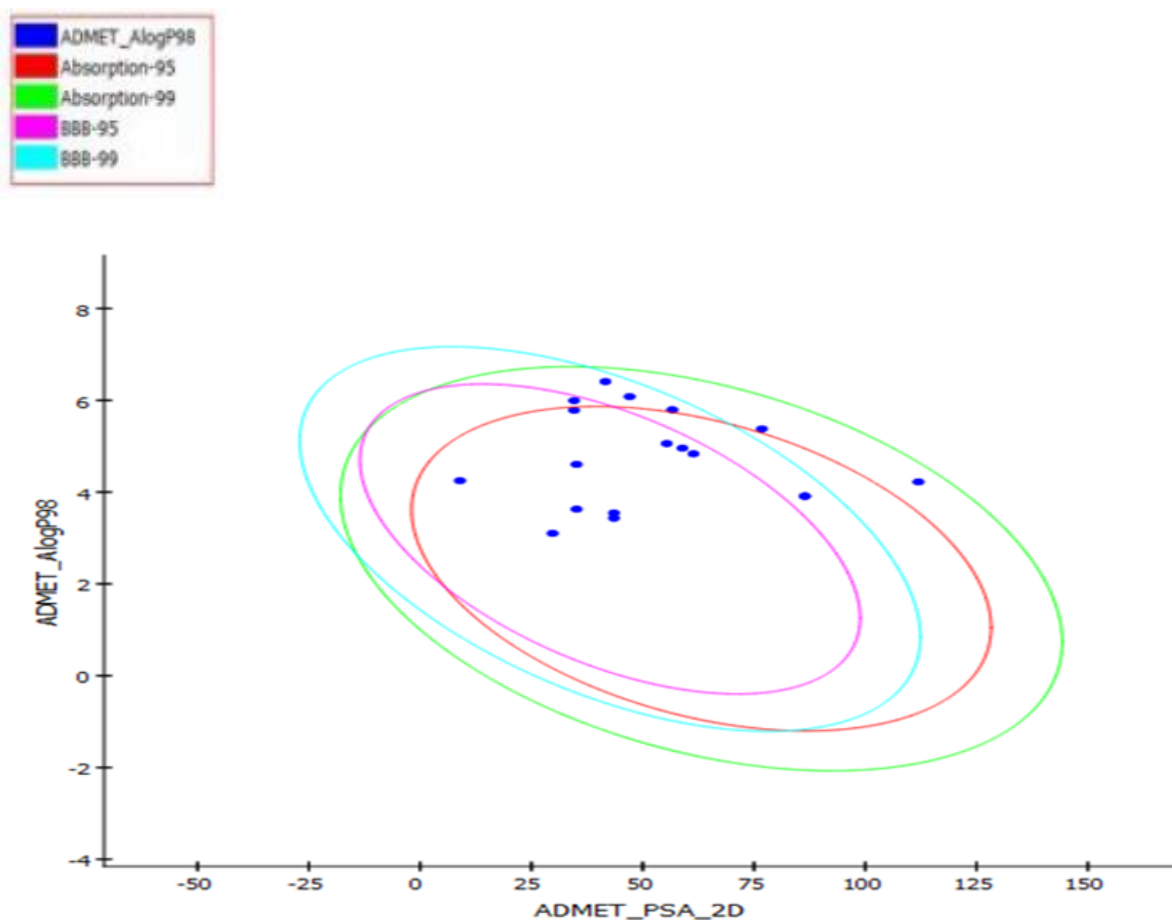
5.9 Insilico ADMET Analysis

Discovery studio 2.0 was used to predict ADMET descriptors of the marine compound

These descriptors include Aqueous solubility of each compound in water at Blood-Brain Barrier penetration (BBB), Cytochrome enzyme inhibition, potential Hepatotoxicity.

It was found that many of the synthesized compounds have PPB level equal to zero or one while the natural compounds have PPB level equal to two or one. This indicates that synthetic compound derivatives have plasma protein binding less than 95%, while natural compounds derivatives have plasma protein binding more than 95%, revealing a high probability that natural compounds to reach the desired targets than the synthetic compounds. The compounds that satisfied the above mentioned properties were selected for molecular docking studies.

Fig 5.5 - ADMET Analysis of Red Algae

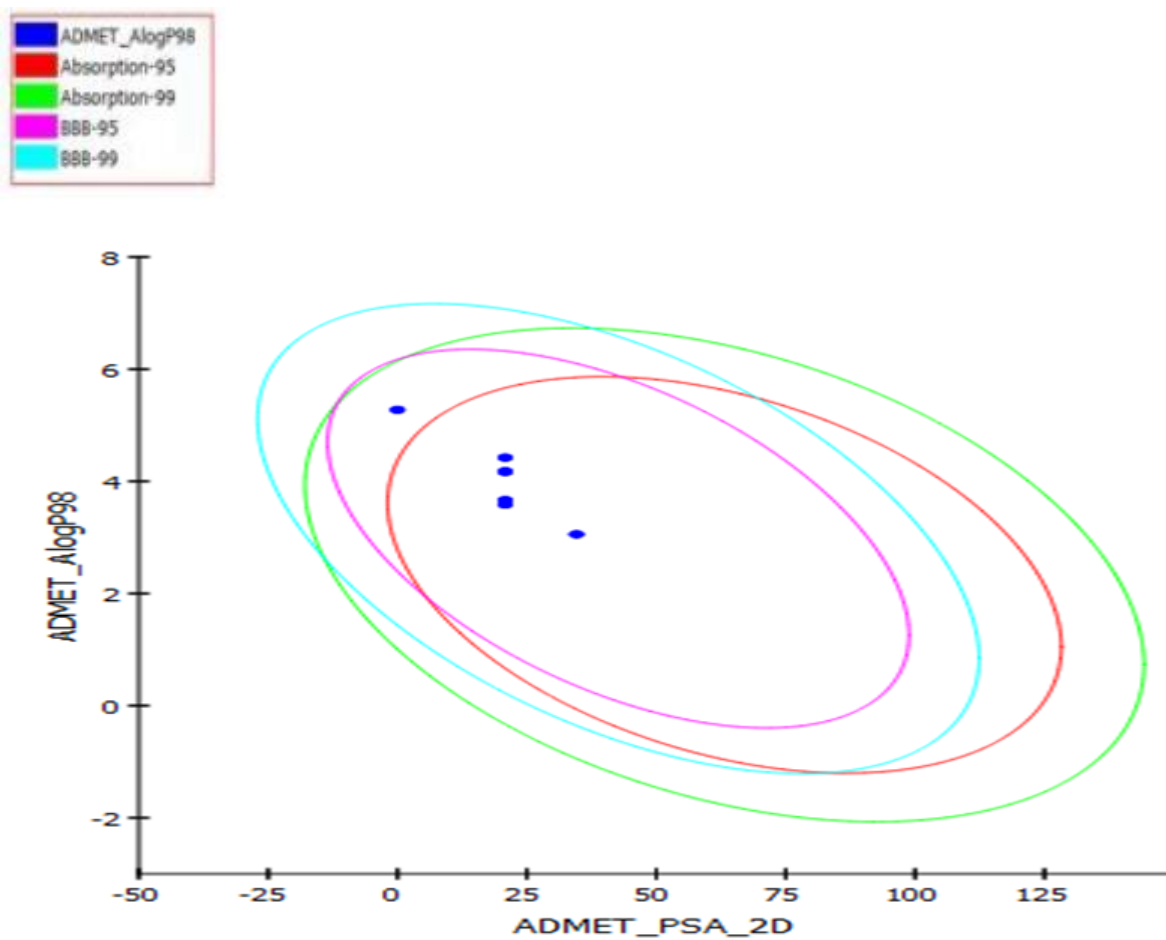


Red colour : Absorption (95%)

Green colour : BBB level (99%)

Blue colour : Absorption (99%)

Fig 5.6-ADMET analysis of Green algae



BBB level: Blood brain barrier level(95%)

Red colour : Absorption (95%)

Green colour : BBB level (99%)

Blue colour : Absorption (99%)

Molecular docking is an effective method that guides us to screen the compounds that accommodate well at the proteins active site and reveals an ideal binding mode of the small molecules, and was used to clarify the binding mode of the compounds to elucidate new information for further structural optimization. The efficacy of a drug also depends on its binding mode and affinity toward the target site, somehow the docking study helps to perceive this efficacy and affinity. LigandFit was implemented in the Discovery Studio software (version 2.0, Accelrys Software) was used for molecular docking studies. After docking the Marine compounds into GAL3 pockets, For each Protein ligand complex, the docked conformation which had the highest docking score with highest score according to scoring function was taken as the best pose among the generated ten poses. The docked conformation which had the highest docking score was selected to analyse the binding pattern because where the higher the values, the greater the favourable binding between the protein and the ligand. The docking studies calculations evaluated the docking behaviour of marine ligands molecules with GAL3.

5.14 DISCUSSION

Cancer is essentially a genetic disease, and many genetic alterations accumulate during the multistep process of carcinogenesis, which eventually leads to abnormal and unrestrained cell growth which ultimately results malignant phenotype (Carter SL et al., 2006).Adenocarcinoma is a kind of cancer that forms in mucus-secreting gland cells, that are found in tissues that line internal organs. When adenocarcinoma spreads from the initial site, it's represented as metastatic (Julie Marks.,2019)The hallmarks of cancer imply abilities of limitless replicative potential, self sufficiency at growth, anti-apoptotic potential, sustained development and also the potential for invasion and metastasis. The 'abilities' are non inheritable due to dysregulation of signaling pathways. For a malignancy to occur, the underlying genetic mechanisms may include gene activation, or via loss of tumour suppressor gene operate these mechanisms are understood to be concerned within the etio-pathogenesis of carcinoma(Swaroop Revannasiddaiah et al.,2014)

In the present study, overlapping DEG's among lung cancer, esophageal cancer, pancreatic, prostate, colorectal, gastric were mined to reveal the common genes, which may indicate a common method of tumorigenesis which ultimately leads to adenocarcinoma. And those genes were further assessed for their hubness to screen the promising biomarker. An integrative approach utilizing differential expression and network analyses was employed for this purpose.

The present study involved differentially expressed at the malignant stage compared with normal tissues indicating that the occurrence and development of adenocarcinoma is closely associated with genetic mutations.

CytoHubba for the identification of the hub genes which are associated with adenocarcinoma. The PPI networks of those overlapping genes were further assessed for hubness based on Degree measures to reveal the top 10 hub gene. Among them top one hub genes we predicted GAL3. Previous studies have reported some of these gene.

Galectin-3 a member of the galectin family is an endogenous β -galactoside binding glycoprotein. it's been found to be related to cell adhesion, recognition, proliferation, differentiation, immunomodulation, growing, cell death and might be a reliable marker for cancer aggressiveness. protoplasm gal-3 immunoexpression appears to be a prognostic about large intestine cancer as a result of the next risk of repetition has been ascertained in tumours with a high score of galectin-3. The expression of gal-3 is related to tumor progression and pathologic process potential in stomach cancers(Thiago Simão Gomes et al., 2014)

CHAPTER-6

SUMMARY AND CONCLUSION

Adenocarcinoma are malignant lumps which are formed from glandular structures in epithelial tissue. They occur from a variety of tissues. These types of cancer account for 80% to 95%. Malignant transformation of adenocarcinoma has been described in the literature as occurring mainly in relationship with the lung, esophageal, colorectal, pancreatic, prostate and gastric . Based on clinical studies showed the lung transformation to malignant lung tumors, and were known as Lung adenocarcinoma(LUAD).The increased risk of lung cancer associated with adenocarcinoma is biologically and pathologically plausible. Statistical analysis suggests that adenocarcinoma was associated with an increased risk of lung tumour, which occurs in both men and women.

CytoHubba which may serve as the promising disease marker and could also serve as a therapeutic target. Minimum 2 fold increased expression of GAL3 is recorded in many types of cancer that includes lung, gastric and esophageal.

Finally, the identified DEG's and Hub genes may help to guide investigations on the mechanism underlying the development and progression of adenocarcinoma associated with above molecular interaction results (i.e) the natural (Marine) (Red algae and Green algae)compounds shown equal suppressing or inhibiting effect of synthetic compounds against Lung,Gastric and Esophageal cancers.

REFERENCES

1. 29 November 2019 — Identification of prognostic gene signature associated with microenvironment of lung adenocarcinoma. Cheng Yue,[#] Hongtao Ma,[#] and Yubai Zhou
2. 13 November 2020-Identification of immune-related gene signature predicting survival in the tumor microenvironment of lung adenocarcinoma- Mengnan Zhao, Ming Li, Zhencong Chen, Yunyi Bian, Yuansheng Zheng, Zhengyang Hu, Jiaqi Liang, Yiwei Huang, Jiacheng Yin, Cheng Zhan, Mingxiang Feng & Qun Wang
3. 4 March 2019- Identification of an immune signature predicting prognosis risk of patients in lung adenocarcinoma-Qian Song, Jun Shang, Zuyi Yang, Lanlin Zhang, Chufan Zhang, Jianing Chen & Xianghua Wu.
4. 23 September 2020 -Identification of a Novel Tumor Microenvironment–Associated Eight-Gene Signature for Prognosis Prediction in Lung Adenocarcinoma- Chao Ma^{1,2*}, Huan Luo^{1,3}, Jing Cao⁴, Xiangyu Zheng⁵, Jinjun Zhang⁵, Yanmin Zhang⁵ and Zongqiang Fu⁵
5. 22 May 2020- Identification of prognostic signature of non–small cell lung cancer based on TCGA methylation data- Yifan Wang, Ying Wang, Ying Wang & Yongjun Zhang.
6. 1 December 2016-Lung cancer: biology and treatment options Hassan Lemjabbar-Alaoui,^a Omer Hassan,^a Yi-Wei Yang,^a and Petra Buchanana
1 Jan 2020, Comprehensive analysis of prognostic immune-related genes in the tumor microenvironment of lung adenocarcinoma-Chao Ma^{1, 2, #, §}, Huan Luo^{1, 3}, Jing Cao^{4, *}

7. 30 Oct 2020-Prognostic Implications of Novel Ten-Gene Signature in Uveal Melanoma-Huan Luo 1 2 3, Chao Ma 2 4, Jinping Shao 1, Jing Cao 1
8. 24 June 2020-Immune profile of the tumor microenvironment and the identification of a four-gene signature for lung adenocarcinoma-Tao Fan^{1,2} , Mingchuang Zhu^{1,2} , Liyu Wang² , Yu Liu² , He Tian² , Yujia Zheng² , Fengwei Tan² , Nan Sun² , Chunxiang Li² , Jie He^{1,2}
9. 1 sep 2015-Esophageal diseases-W. Asher Wolf, Evan S. Dellon, and Nicholas J. Shaheen
- 10.21 Nov 2018-Pancreatic cancer: A review of clinical diagnosis, epidemiology, treatment and outcomes-Andrew McGuigan, Paul Kelly, Richard C Turkington, Claire Jones, Helen G Coleman, and R Stephen McCain
- 11.1 Oct 2015-Prevention and Early Detection of Prostate Cancer-Jack Cuzick et al
- 12.6 Oct 2017-Colorectal cancer-Anne B Ballinger
- 13.13 Nov 2003-Cytoscape: A Software Environment for Integrated Models of Biomolecular Interaction Networks-Paul Shannon et al.
- 14.1 Jan 2020-*In silico* study for evaluating the binding mode and interaction of 1, 2, 4-triazole and its derivatives as potent inhibitors against Lipoate protein B (LipB)-Shola ElijahAdenijiSaniUbaAdamuUzairu
- 15.19 Oct 2018-Galectins in Cancer and Translational Medicine: From Bench to Bedside-Armando Bartolazzi