**ISLAMIC UNIVERSITY OF TECHNOLOGY**

UNDER GRADUATE THESIS

# Gene Co-expression Network analysis of Lung Adenocarcinoma Cell Carcinoma data


Authors

Md. Saif Uddin (124449)

Md. Tanvir Ahamed (124444)

*Supervisor*

Tareque Mohmud Chowdhury

Assistant Professor

Department of Computer Science and Engineering


A thesis submitted to the Department of CSE in fulfilment of the requirements for the Degree of B.Sc Engineering in CSE.

Academic Year: 2015-16

**November 2016**

# *Declaration of Authorship*

We,Md. Saif Uddin (124449), Md. Tanvir Ahamed (124444), declare that this thesis titled, "Gene Co-expression Network analysis of Lung Adenocarcinoma Cell Carcinoma data" and the work presented in it are our own. We confirm that this work was done wholly or mainly while in candidature for a research degree at this University. Where any part of this thesis has previously been submitted for a degree or any other qualification at this University or any other institution, this has been clearly stated. In which place we have consulted the published work of others, this is always clearly attributed. If we have quoted from the work of others, the source is always given. With the exception of such quotations, this thesis is entirely our own work. We have acknowledged all main sources of help. Where the thesis is based on work done by ourselves jointly with others, we have made clear exactly what was done by others and what we have contributed ourselves.

Authors:

---

Md. Saif Uddin
Student ID:124449

---

Md. Tanvir Ahamed
Student ID:124444

Supervisor:

---

Tareque Mohmud Chowdhury
Assistant Professor
CSE, IUT.

## Abstract

A gene co-expression analysis on Lung adenocarcinoma data was done to find modules of genes that might highly impact the growth of this type of tumor. Along with that, cancer survival data was used to relate modules to prognostic significance for survival time. Analysis on microarray data revealed modules that were significant in gene enrichment analysis and 4 genes - TTk, C6orf173, CENPE, DCC1 were found that were significant in terms of survival time. A second analysis was done on a second set of RNAseq data and the significant genes in modules was found there, were also found in the RNAseq data implying that these genes might indeed play a crucial role in Lung adenocarcinoma.

# Table of Contents


1.1 Overview ..... 8

1.3 Research Challenges ..... 9

1) Analyzing & Understanding the whole methodology clearly. ..... 9

2) Examined the previous works on this methodology. ..... 9

3) Collecting The Dataset Which we want to work on. ..... 9

4) Works with a new statistical software **R**. ..... 9

5) Detection of hub genes. ..... 9

6) Module Construction ..... 9

1.4 Motivation ..... 9

1.5 Objectives ..... 10

2.1 LUNG CANCER ..... 11

2.2 The lungs and breathing system ..... 12

2.2.1 Types of Lung Cancer ..... 13

2.3 Network Analytical View ..... 14

2.3.1 Gene Co-expression Network ..... 14

2.3.2 Gene Regulatory Network ..... 15

2.4 Related Works for Gene Co-expression Network ..... 17

2.4.1 Analysis of oncogenic signaling networks in glioblastoma identifies ASPM as a molecular target ..... 17

2.4.2 Correlating transcriptional networks to breast cancer survival: a large-scale co-expression analysis ..... 19

2.4.3 Gene Networks and microRNAs Implicated in Aggressive Prostate Cancer ..... 19

3.1 Overall Concept ..... 21

3.2 Proposed Method ..... 21

3.3 Proposed Method Overview ..... 22

3.4 Significant Gene Finding: ..... 26

3.5 Experimental Data and Results: ..... 27

3.5.1Datasets: ..... 27

3.4.3 Result of Simulation: ..... 38
3.4.4 Comparisons between two dataset: ..... 39
3.4.5 Result Verification: ..... 40

4. Summary of the research ..... 41
5. References ..... 39

# Chapter 1

# Introduction

## 1.1 Overview

Cancer is a group of diseases involving abnormal cell growth with the potential to invade or spread to other parts of the body. Not all tumors are cancerous; benign tumors do not spread to other parts of the body. Possible signs and symptoms include a lump, abnormal bleeding, prolonged cough, unexplained weight loss and a change in bowel movements. While these symptoms may indicate cancer, they may have other causes. Over 100 cancers affect humans. In 2012 about 14.1 million new cases of cancer occurred globally (not including skin cancer other than melanoma). It caused about 8.2 million deaths or 14.6% of human deaths. ***The most common types of cancer in males are lung cancer, prostate cancer, colorectal cancer and stomach cancer***. In females, the most common types are breast cancer, colorectal cancer, lung cancer and cervical cancer. If skin cancer other than melanoma were included in total new cancers each year it would account for around 40% of cases. **Cancer** form by neoplasms. A neoplasm or tumor is a group of cells that have undergone unregulated growth and will often form a mass or lump, but may be distributed diffusely. These characteristics are required to produce a malignant tumor because all tumor are not malignant to occur cancer.

- Cell growth
- Limitless growth in case.
- Cell Death
- Cell Division in exponential rate.
- Invasion of tissue and formation of metastases.

## 1.2 Problem Statement

Actually most cancers do not cause any symptoms until they have spread, but some people with early lung cancer do have symptoms. We will detect the early stage lung cancer. The Weighted Gene Co-expression Network Analysis (WGCNA) is the method to be used.

## 1.3 Research Challenges

1) Analyzing & Understanding the whole methodology clearly.

2) Examined the previous works on this methodology.

3) Collecting The Dataset Which we want to work on.

4) Works with a new statistical software **R**.

5) Detection of hub genes.

6) Module Construction

## 1.4 Motivation

At first a question may arise that why we choose lung cancer? why not the others? The main reason is if we have a look on an early statistics of different types of cancer then we see that the cancer occurrence of prostate cancer is the highest one 25% and lung cancer is very close to it near 15% which rate is pretty higher than others. On the contrary, the cancer mortality rate of lung cancer is higher than others.

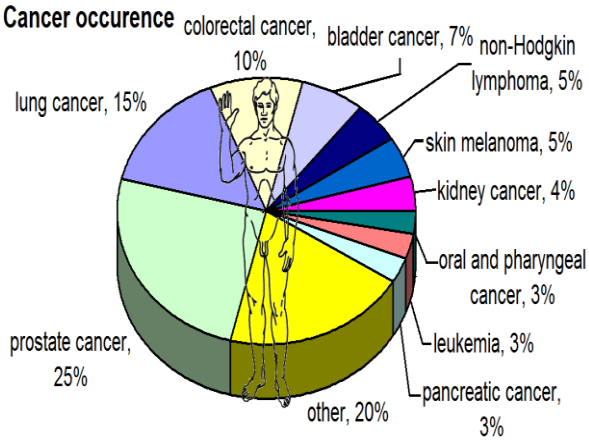



Fig 1: Cancer Occurrence

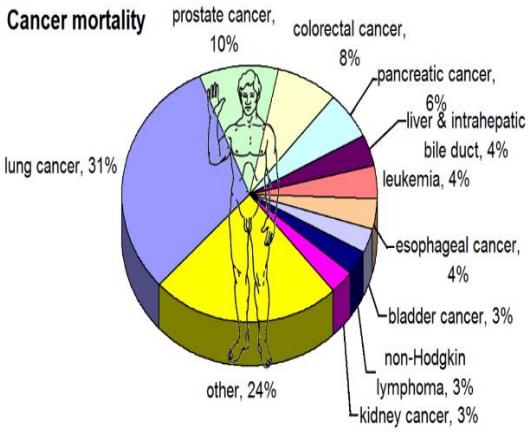



Fig 2: Cancer Mortality

There are many way to detect lung cancer but there are few ways to determine why it occurred. So far the results are still pending, these scans is the best way of finding lung cancer before it spreads. We want to detect Lung cancer accurately by an another method using Gene Co-expression Network Weighted Gene Co-expression Network Analysis (WGCNA).

## 1.5 Objectives

Our main objective of this thesis is to verify a method that will analyze the microarray data of Leukemia (cancer of the blood cells) sampled from different races. Our main objective is to-

- Identifying cancer affected gene
- Verification of this method
- Drug identification for specific cancer.

Already we have observe different types of cancer detection using Weighted Gene Co-expression Network Analysis (WGCNA) method .Now, Our goal is to ensure how efficiently we can detect the affected genes for leukemia.

# Chapter 2

# Literature Reviews

This chapter is divided into two categories. In the first part of this chapter, we will describe the reason of lung cancer. It gives the basics idea about the lung cancer's occurrences and the detection made latter part of this book. Information about cancer detection and other relevant topics are mostly taken from various research articles [2]. In the last section, a short description about the different properties of the method of lung cancer detection has been demonstrated.

## 2.1 LUNG CANCER

Lung cancer can start in the windpipe (trachea), the main airway (bronchus) or the lung tissue. Find out about symptoms, risk factors and causes of lung cancer, diagnostic tests, treatment, including surgery, chemotherapy, radiotherapy and biological therapy, likely outcome (prognosis), research and how to cope with lung cancer, including managing breathlessness.

By far the biggest cause of lung cancer is smoking. It causes more than 8 out of 10 cases (86%) including a small proportion caused by exposure to second hand smoke in nonsmokers (passive smoking).

Here are some facts about smoking and lung cancer:

- The more you smoke, the more likely you are to get lung cancer **but** the length of time you have been a smoker is even more important than how many cigarettes you smoke a day
- Starting smoking at a young age is even more harmful than starting as an adult
- Stopping smoking reduces your risk of lung cancer compared to continuing to smoke. The sooner you quit, the better your health - but it's never too late
- Passive smoking (breathing in other people's cigarette smoke) increases the risk of lung cancer, but it is still much less than if you smoke yourself

## 2.2 The lungs and breathing system

These are part of the body system we use to breathe - the respiratory system. It is made up of the

- Nose and mouth
- Windpipe (trachea)
- Airways to each lung (the right main bronchus and left main bronchus)

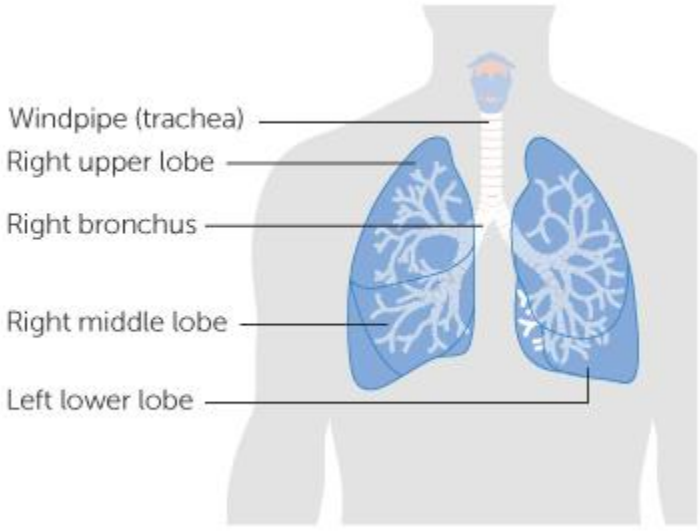



Fig 3: Lung and breathing system

The lungs bring oxygen into our bodies and pass it into the bloodstream so that it can circulate to every body cell. We use the muscles of our chest and a large flat muscle under the lungs (the diaphragm – pronounced di-a-fram) to draw air into the lungs. The diaphragm is at the base of the chest cavity, just above the stomach. The chest cavity is sealed so that when you breathe in and the muscles make it bigger, this creates a vacuum inside, which draws air in through your nose and down into the lungs.

At the end of the smallest airways in the lungs (the secondary bronchii) are the smallest tubes, the bronchioles. These carry air throughout the lungs. At the end of the bronchioles are air sacs called alveoli. There are millions of these tiny sacs. This is where oxygen is absorbed into the bloodstream from the air that we breathe in. Once in the blood, the oxygen can travel throughout the body and reach every body cell.

As oxygen is being absorbed, carbon dioxide passes back into the alveoli from the bloodstream. This waste gas is removed from the body as we breathe out.

They are taking the challenges to find out the Motif sequences which have great impact on biological function. Motif identification can be thought of as finding the best local multiple alignments for the sequences under consideration. But there are some causes for what it is becoming challenging and difficult task for the bioinformatics people.

### 2.2.1 Types of Lung Cancer

Lung Cancer is the event of tumors in lung where uncontrolled growth of tissues happen. Lung Cancer is mainly of 2 types.

- **Non-small cell lung carcinoma (NSCLC) and**
- **Small cell lung carcinoma(SCLC)**.

**Non-small-cell lung carcinoma** (**NSCLC**) is any type of epithelial lung cancer other than small cell lung carcinoma (SCLC). NSCLC accounts for about 85% of all lung cancers. As a class, NSCLCs are relatively insensitive to chemotherapy, compared to small cell carcinoma. When possible, they are primarily treated by surgical resection with curative intent, although chemotherapy is increasingly being used both pre-operatively (neoadjuvant chemotherapy) and post-operatively (adjuvant chemotherapy). The most common types of NSCLC are :

- **Squamous cell carcinoma**
- **Large cell carcinoma**
- **Adenocarcinoma**

**Adenocarcinoma** of the lung is currently the most common type of lung cancer in "never smokers" (lifelong non-smokers). Adenocarcinomas account for approximately 40% of lung

cancers. Historically, adenocarcinoma was more often seen peripherally in the lungs than small cell lung cancer and squamous cell lung cancer, both of which tended to be more often centrally located. However, recent studies suggest that the "ratio of centrally-to-peripherally occurring" lesions may be converging toward unity for both adenocarcinoma and squamous cell carcinoma.

**Squamous cell carcinoma** (SCC) of the lung is more common in men than in women. It is closely correlated with a history of tobacco smoking, more so than most other types of lung cancer. According to the Nurses' Health Study, the relative risk of SCC is approximately 5.5, both among those with a previous duration of smoking of 1 to 20 years, and those with 20 to 30 years, compared to never-smokers.[11] The relative risk increases to approximately 16 with a previous smoking duration of 30 to 40 years, and approximately 22 with more than 40 years.

**Large cell lung carcinoma** (LCLC) is a heterogeneous group of undifferentiated malignant neoplasms originating from transformed epithelial cells in the lung. LCLC's have typically comprised around 10% of all NSCLC in the past, although newer diagnostic techniques seem to be reducing the incidence of diagnosis of "classic" LCLC in favor of more poorly differentiated squamous cell carcinomas and adenocarcinomas. LCLC is, in effect, a "diagnosis of exclusion", in that the tumor cells lack light microscopic characteristics that would classify the neoplasm as a small-cell carcinoma, squamous-cell carcinoma, adenocarcinoma, or other more specific histologic type of lung cancer. LCLC is differentiated from small cell lung carcinoma (SCLC) primarily by the larger size of the anaplastic cells, a higher cytoplasmic-to-nuclear size ratio, and a lack of "salt-and-pepper" chromatin.

## 2.3 Network Analytical View

### 2.3.1 Gene Co-expression Network

A **gene co-expression network (GCN)** is an undirected graph, where each node corresponds to a gene, and a pair of nodes is connected with an edge if there is a significant co expresson relationship between them. Having gene expression profiles of a number of genes for several samples or experimental conditions, a gene co-expression network can be constructed by looking for pairs of genes which show a similar expression pattern across samples, since the transcript levels of two co-expressed genes rise and fall together across samples. Gene co-expression networks are of biological interest since co-expressed genes are controlled by the same transcriptional regulatory program, functionally related, or members of the same pathway or protein complex.

The direction and type of co-expression relationships are not determined in gene co-expression networks; whereas in a gene regulatory network (GRN) a directed edge connects two genes, representing a biochemical process such as a reaction, transformation, interaction, activation or

inhibition. Compared to a GRN, a GCN does not attempt to infer the causality relationships between genes and in a GCN the edges represent only a correlation or dependency relationship among genes. Modules or the highly connected subgraphs in gene co-expression networks correspond to clusters of genes that have a similar function or involve in a common biological process which causes many interactions among themselves.

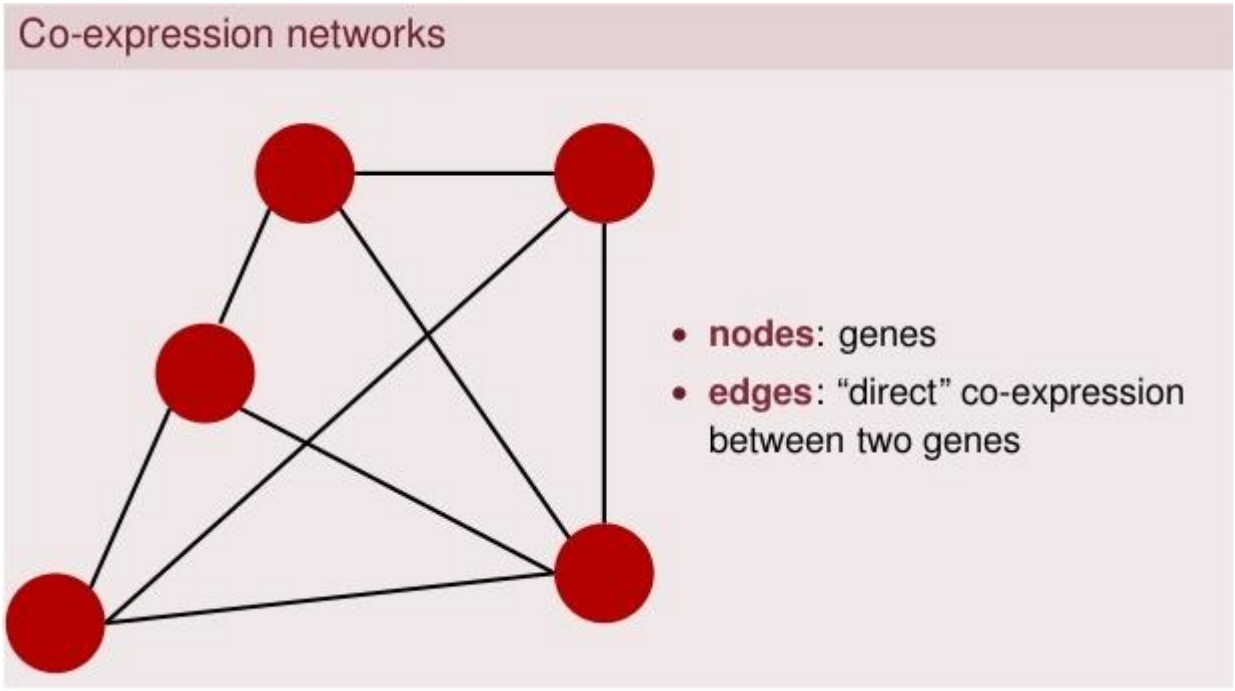



**Figure 4:** Gene Co-expression Network.

### 2.3.2 Gene Regulatory Network

A **gene (or genetic) regulatory network (GRN)** is a collection of molecular regulators that interact with each other and with other substances in the cell to govern the gene expression levels of mRNA and proteins. These play a central role in morphogenesis, the creation of body structures, which in turn is central to evolutionary developmental biology (evo-devo).

The regulator can be DNA, RNA, protein and complexes of these. The interaction can be direct or indirect (through transcribed RNA or translated protein). In general, each mRNA molecule goes on to make a specific protein (or set of proteins). In some cases this protein will be structural, and will accumulate at the cell membrane or within the cell to give it particular structural properties. In other cases the protein will be an enzyme, i.e., a micro-machine that catalyses a certain reaction, such as the breakdown of a food source or toxin. Some proteins though serve only to activate other genes, and these are the transcription factors that are the main players in regulatory networks or cascades. By binding to the promoter region at the start of other genes they turn them on, initiating the production of another protein, and so on. Some transcription factors are inhibitory.In single-celled organisms, regulatory networks respond to the

external environment, optimising the cell at a given time for survival in this environment. Thus a yeast cell, finding itself in a sugar solution, will turn on genes to make enzymes that process the sugar to alcohol.[1] This process, which we associate with wine-making, is how the yeast cell makes its living, gaining energy to multiply, which under normal circumstances would enhance its survival prospects.

In multicellular animals the same principle has been put in the service of gene cascades that control body-shape.[2] Each time a cell divides, two cells result which, although they contain the same genome in full, can differ in which genes are turned on and making proteins. Sometimes a 'self-sustaining feedback loop' ensures that a cell maintains its identity and passes it on. Less understood is the mechanism of epigenetics by which chromatin modification may provide cellular memory by blocking or allowing transcription. A major feature of multicellular animals is the use of morphogen gradients, which in effect provide a positioning system that tells a cell where in the body it is, and hence what sort of cell to become. A gene that is turned on in one cell may make a product that leaves the cell and diffuses through adjacent cells, entering them and turning on genes only when it is present above a certain threshold level. These cells are thus induced into a new fate, and may even generate other morphogens that signal back to the original cell. Over longer distances morphogens may use the active process of signal transduction. Such signalling controls embryogenesis, the building of a bodyplan from scratch through a series of sequential steps. They also control and maintain adult bodies through feedback processes, and the loss of such feedback because of a mutation can be responsible for the cell proliferation that is seen in cancer. In parallel with this process of building structure, the gene cascade turns on genes that make structura proteins that give each cell the physical properties it needs.

.

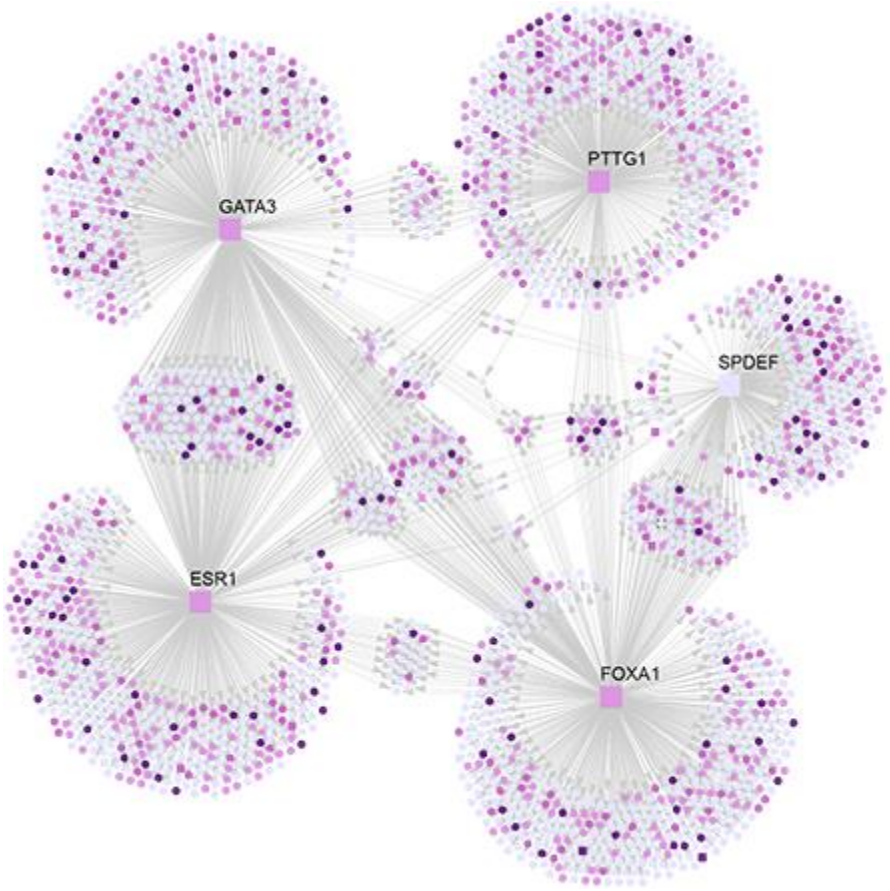



**Figure 4**. Gene Regulatory Network

## 2.4 Related Works for Gene Co-expression Network

In recent years, there is an incredible growth of network methods to explore functionality of genes from the system level. However, most researchers are focused on the un-weighted networks.A new weighted networks method is established. To evaluate the efficiency of the method, application of the method to screen diabetic candidate pathogenic gene and module recognition are also given

### 2.4.1 Analysis of oncogenic signaling networks in glioblastoma identifies ASPM as a molecular target

Glioblastoma is the most common primary malignant brain tumor of adults and one of the most lethal of all cancers. Patients with this disease have a median survival of 15 months from the time of diagnosis despite surgery, radiation, and chemotherapy. New treatment approaches are needed. Recent works suggest that glioblastoma patients may benefit from molecularly targeted

therapies. Here, we address the compelling need for identification of new molecular targets. Leveraging global gene expression data from two independent sets of clinical tumor samples (n=55 and n=65), they identify a gene co-expression module in glioblastoma that is also present in breast cancer and significantly overlaps with the ‘‘metasignature’’ for undifferentiated cancer. Studies in an isogenic model system demonstrate that this module is downstream of the mutant epidermal growth factor receptor, EGFRvIII, and that it can be inhibited by the epidermal growth factor receptor tyrosine kinase inhibitor Erlotinib. They identify ASPM (abnormal spindle-like microcephaly associated) as a key gene within this module and demonstrate its overexpression in glioblastoma relative to normal brain (or body tissues). Finally, they show that ASPM inhibition by siRNA-mediated knockdown inhibits tumor cell proliferation and neural stem cell proliferation, supporting ASPM as a potential molecular target in glioblastoma [2].

Main Objective:

Their weighted gene co-expression network analysis provides a blueprint for leveraging genomic data to identify key control networks and molecular targets for glioblastoma.

Process:

Firstly, Microarray Data. Glioblastoma gene expression profiling with Affymetrix high-density oligonucleotide microarrays was performed and analyzed. Quantification was performed by using model-based expression and the perfect match minus mismatch method implemented in dCHIP. We used the breast cancer microarray data. In brief, cell lines were grown in duplicate cultures under serum free conditions for 48 h, and RNA was isolated by using the Qiagen (Valencia, CA) RNeasy Mini Kit Gene. Expression analysis by using Affymetrix HG-U133A arrays was performed and analyzed, as described above.

EGFR Inhibitor Treatment and siRNA Studies. The EGFR tyrosine kinase inhibitor Erlotinib (Tarceva, OSI-774) was kindly provided by Genentech (South San Francisco, CA). U87MG and U87-EGFRvIII cells (1 105) were seeded, respectively, in 100-mm culture dishes and maintained in DMEM supplemented with 10% FBS. Cells were incubated in 5% CO2, 95% humidity incubator for 3 days to reach 50–70% confluency. Then all cells were switched to serum-free medium. The next day U87-EGFRvIII cells were treated by 5M OSI-774, whereas U87MG and U87-EGFRvIII control group received the equivalent vehicle. Twenty-four hours later, cell total RNA was isolated by Qiagen RNeasy Mini Kit. RT-PCR analysis of expression of selected genes after treatment is described in the Supporting Methods, which is published as supporting information on the PNAS web site. The specific methods for siRNA studies are available in Supporting Methods. For proliferation assays, 1,500 cells per well in eight replicates were seeded into 96-well plates. Cells were fixed and stained by 0.25% crystal violet in methanol every day or every other day.

### 2.4.2 Correlating transcriptional networks to breast cancer survival: a large-scale co-expression analysis

They have utilized WGCNA to identify 11 coregulated gene clusters across 2342 breast cancer samples from 13 microarray-based gene expression studies. A number of these transcriptional modules were found to be correlated to clinic pathological variables (e.g. tumor grade), survival endpoints for breast cancer as a whole (disease-free survival, distant disease-free survival and overall survival) and also its molecular subtypes (luminal A, luminal B, HER2+ and basal-like). Examples of findings arising from this work include the identification of a cluster of proliferation-related genes that when upregulated correlated to increased tumor grade and were associated with poor survival in general.

### 2.4.3 Gene Networks and microRNAs Implicated in Aggressive Prostate Cancer

Prostate cancer, a complex disease, can be relatively harmless or extremely aggressive. To identify candidate genes involved in causal pathways of aggressive prostate cancer, we implemented a systems biology approach by combining differential expression analysis and coexpression network analysis to evaluate transcriptional profiles using lymphoblastoid cell lines from 62 prostate cancer patients with aggressive phenotype (Gleason grade $\geq$ 8) and 63 prostate cancer patients with nonaggressive phenotype (Gleason grade $\leq$ 5). From 13,935 mRNA genes and 273 microRNAs (miRNA) tested, we identified significant differences in 1,100 mRNAs and 7 miRNAs with a false discovery rate (FDR) of <0.01. We also identified a coexpression module demonstrating significant association with the aggressive phenotype of prostate cancer ($P = 3.67 \times 10-11$). The module of interest was characterized by overrepresentation of cell cycle–related genes (FDR = $3.50 \times 10-50$). From this module, we further defined 20 hub genes that were highly connected to other genes. Interestingly, 5 of the 7 differentially expressed miRNAs have been implicated in cell cycle regulation and 2 (miR-145 and miR-331-3p) are predicted to target 3 of the 20 hub genes. Ectopic expression of these two miRNAs reduced expression of target hub genes and subsequently resulted in cell growth inhibition and apoptosis. These results suggest that cell cycle is likely to be a molecular pathway causing aggressive phenotype of prostate cancer. Further characterization of cell cycle–related genes (particularly, the hub genes) and miRNAs that regulate these hub genes could facilitate identification of candidate genes responsible for the aggressive phenotype and lead to a better understanding of prostate cancer etiology and progression. [Cancer Res 2009;69(24):9490–7]

# Chapter 3

# Proposed Method

## 3.1 Overall Concept

Correlation networks are increasingly being used in bioinformatics applications. For example, weighted gene co-expression network analysis is a systems biology method for describing the correlation patterns among genes across microarray samples. Weighted correlation network analysis (WGCNA) can be used for finding clusters (modules) of highly correlated genes, for summarizing such clusters using the module eigengene or an intramodular hub gene, for relating modules to one another and to external sample traits (using eigengene network methodology), and for calculating module membership measures. Correlation networks facilitate network based gene screening methods that can be used to identify candidate biomarkers or therapeutic targets. These methods have been successfully applied in various biological contexts, e.g. cancer, mouse genetics, yeast genetics, and analysis of brain imaging data. While parts of the correlation network methodology have been described in separate publications, there is a need to provide a userfriendly, comprehensive, and consistent software implementation and an accompanying tutorial.

## 3.2 Proposed Method

The WGCNA package contains a comprehensive set of functions for performing a correlation network analysis of large, high-dimensional data sets. Functions in the WGCNA package can be divided into the following categories:

- Network construction;
- Module detection;
- Module and gene selection;
- Calculations of topological properties;
- Data simulation;
- Visualization;
- Interfacing with external software packages.

An exhaustive list of implemented functions together with detailed descriptions is provided in the R package manual. Here we briefly outline the main functionality of the package.

## 3.3 Proposed Method Overview

### Category 1: Functions for network construction :

A network is fully specified by its adjacency matrix aij, a symmetric n × n matrix with entries in [0, 1] whose component $a_{ij}$ encodes the network connection strength between nodes i and j. To calculate the adjacency matrix, An intermediate quantity called the co-expression similarity $S_{ij}$ is first defined. The default method defines the co-expression similarity $S_{ij}$ as the absolute value of the correlation coefficient between the profiles of nodes i and j:

$$S_{ij} = |cor(x_i, x_j)|.$$

The WGCNA package also implements alternative co-expression measures, e.g. more robust measures of correlation (the biweight mid correlation or the Spearman correlation). A signed co-expression measure can be defined to keep track of the sign of the co-expression information. For convenience, we define the co-expression similarity measure such that it takes on values in [0,1]. Using a thresholding procedure, the co-expression similarity is transformed into the adjacency. An unweighted network adjacency aij between gene expression profiles xiand xj can be defined by hard thresholding the co-expression similarity sij as where $\tau$ is the "hard" threshold parameter. Thus, two genes are linked (aij = 1) if the absolute correlation between their expression profiles exceeds the (hard) threshold $\tau$. The hard-thresholding procedure is implemented in the function signumAdjacencyFunction. While unweighted networks are widely used, they do not reflect the continuous nature of the underlying co-expression information and may thus lead to an information loss. In contrast, weighted networks allow the adjacency to take on continuous values between 0 and 1. A weighed network adjacency can be defined by raising the co-expression similarity to a power [5,10]: with $\beta \geq 1$. The function adjacency calculates the adjacency matrix from expression data. The adjacency in Equation 5 implies that the weighted adjacency aij between two genes is proportional to their similarity on a logarithmic scale, $\log(a_{ij}) = \beta \times \log(s_{ij})$. Adjacency functions for both weighted and unweighted networks require the user to choose threshold parameters, for example by applying the approximate scale-free topology criterion [5]. The package

provides functions pickSoftThreshold, pickHardThreshold that assist in choosing the parameters, as well as the function scaleFreePlot for evaluating whether the network exhibits a scale free topology. Figure 2A shows a plot identifying scale free topology in simulated expression data.

## Category 2: Functions for module detection

Once the network has been constructed, module detection is often a logical next step. Modules are defined as clusters of densely interconnected genes. Several measures of network interconnectedness are described in [4]. As default, we we use the topological overlap measure [4-5] since it has worked well in several applications. WGCNA identifies gene modules using unsupervised clustering, i.e. without the use of a priori defined gene sets. The user has a choice of several module detection methods. The default method is hierarchical clustering using the standard R function hclust [28]; branches of the hierarchical clustering dendrogram correspond to modulesand can be identified using one of a number of available branch cutting methods, for example the constant-height cut or two Dynamic Branch Cut methods [6].

## Category 3: Functions for module and gene selection

Finding biologically or clinically significant modules and genes is a major goal of many co-expression analyses. The definition of biological or clinical significance depends on the research question under consideration. Abstractly speaking, we define a gene significance measure as a function GS that assigns a non-negative number to each gene; the higher GSi the more biologically significant is gene i. In functional enrichment analysis, a gene significance measure could indicate pathway membership. In gene knockout experiments, gene significance could indicate knockout essentiality. A microarray sample trait T can be used to define a trait-based gene significance measure as the absolute correlation between the trait and the expression profiles, Equation 2. A measure of module significance can be defined as average gene significance across the module genes (Figure 3A). When dealing with a sample trait T, a measure of statistical significance between the module eigengene E and the trait T can be defined, for example,using correlation (Equation 2) or a p-value (Equation 3)obtained from a univariate regression model between E and T. Modules with high trait significance may represent pathways associated with the sample trait. Genes with high module membership in modules related to traits (Figure 3B) are natural candidates for further validation[2,7,8,9].

## Category 4: Functions for studying topological properties

Many topological properties of networks can be succinctly described using network concepts, also known as network statistics or indices [11,33]. Network concepts include whole network connectivity (degree), intramodular connectivity, topological overlap, the clustering coefficient,density etc. Differential analysis of network concepts such as intramodular connectivity may reveal regulatory changes in gene expressions [15,18]. The WGCNA package implements several functions, such as softConnectivity, intramodularConnectivity,

TOMSimilarity,clusterCoef, networkConcepts, for computing these network concepts. Basic R functions can be used to create summary statistics of these concepts and for testing their differences across networks.Network concepts for measuring cluster structureGene clustering

$$ClusterCoef_i = \frac{\sum_{l \neq i} \sum_{m \neq i, l} a_{il} a_{lm} a_{mi}}{\left\{ \left( \sum_{l \neq i} a_{il} \right)^2 - \sum_{l \neq i} (a_{il})^2 \right\}}.$$

trees and TOM plots that visualize interconnectivity patterns often suggest the presence of large modules. Network theory offers a wealth of intuitive network concepts for describing the pairwise relationships among genes that are depicted in cluster trees and heat maps [11]. To illustrate this point, we describe two network concepts in the following. By visual inspection of Figures 2C and 4B, genes appear to be highly interconnected, e.g. turquoise module genes form a reddish square in the TOM plot. This property of dense connections among the genes of module q can be measured using the concept of module density, which is defined as the average adjacency of the module genes:where A(q) denotes the n(q) × n(q) adjacency matrix

$$Density(A^{(q)}) = \frac{\sum_i \sum_{j \neq i} a_{ij}^{(q)}}{n^{(q)}(n^{(q)} - 1)}$$

corresponding to the sub-network formed by the genes of module q. Another useful concept is the clustering coefficient of gene i, which is a measure of 'cliquishness' [4]. Specifically,In unweighted networks, ClusterCoefi equals 1 if and only if all neighbors of gene i are also linked to each other. For weighted networks, $0 \leq a_{ij} \leq 1$ implies that $0 \leq ClusterCoef_i \leq 1$ [5]. The mean clustering coefficient has been used to measure the extent of module structure present in a network [10].

## Category 5: Functions for simulating microarray data with modular structure

Simple yet sufficiently realistic simulated data is often important for evaluation of novel data mining methods.The WGCNA package includes simulation functions sim ulateDatExpr, simulateMultiExpr, simulateDatExpr5Modules that result in expression data sets with a customizable modular (cluster) structure. The user can choose the modular structure by specifying a set of seed eigengenes, one for each module, around which each module is built. Module genes are simulated to exhibit progressively lower correlations with the seed which leads to genes with progressively lower intramodular connectivity. The user can specify module sizes and the number of background genes, i.e. genes outside of the modules. The seed eigengenes can be simulated to reflect dependence relationships between the modules (function simulateEigengeneNetwork).

## Category 6: Visualization functions

Module structure and network connections in the expression data can be visualized in several different ways. Forexample, the co-expression module structure can be visualized by heatmap plots of gene-gene connectivity thatcan be produced using the function TOMplot. Examples are presented in Figures 2C and 4B. An alternative is a multi-dimensional scaling plot; an example is presented in Figure 2B. Relationships among modules can be summarized by a hierarchical clustering dendrogram of their eigengenes, or by a heatmap plot of the corresponding eigengene network (function labeledHeatmap), illus trated in Figures 3C, D, and 4C, D. The package includes several additional functions designed to aid the user in visualizing input data and results. These functions rely on basic plotting functions provided in R and the packages sma [11] and fields [12].

## Category 7: Functions for interfacing with other software packages

To enhance the integration of WGCNA results with other network visualization packages and gene ontology analysis software, we have created several R functions and corresponding tutorials. For example, our R functions exportNetworkToVisANT and exportNetworkToCyto

scape allow the user to export networks in a format suitable for VisANT [13] and Cytoscape [14], respectively.

## 3.4 Significant Gene Finding:

A gene co-expression analysis on Lung adenocarcinoma data was done to find modules of genes that might highly impact the growth of this type of tumor. Along with that, cancer survival data was used to relate modules to prognostic significance for survival time. Analysis on microarray data revealed modules that were significant in gene enrichment analysis and 4 genes - TTk, C6orf173, CENPE, DCC1 were found that were significant in terms of survival time. A second analysis was done on a second set of RNAseq data and the significant genes in modules was not found there, were not also found in the RNAseq data implying that these genes might indeed play a crucial role in Lung adenocarcinoma.

## 3.5 Experimental Data and Results:

In this paper, we have used the gene expression microarray and gene expression RNAseq datasets. The data was taken from TCGA data portal. The number of genes in microarray is 17,815 and the number of genes of RNAseq is 20.531.

### 3.5.1 Datasets:

The datasets that we used are as follows (accession numbers specified)

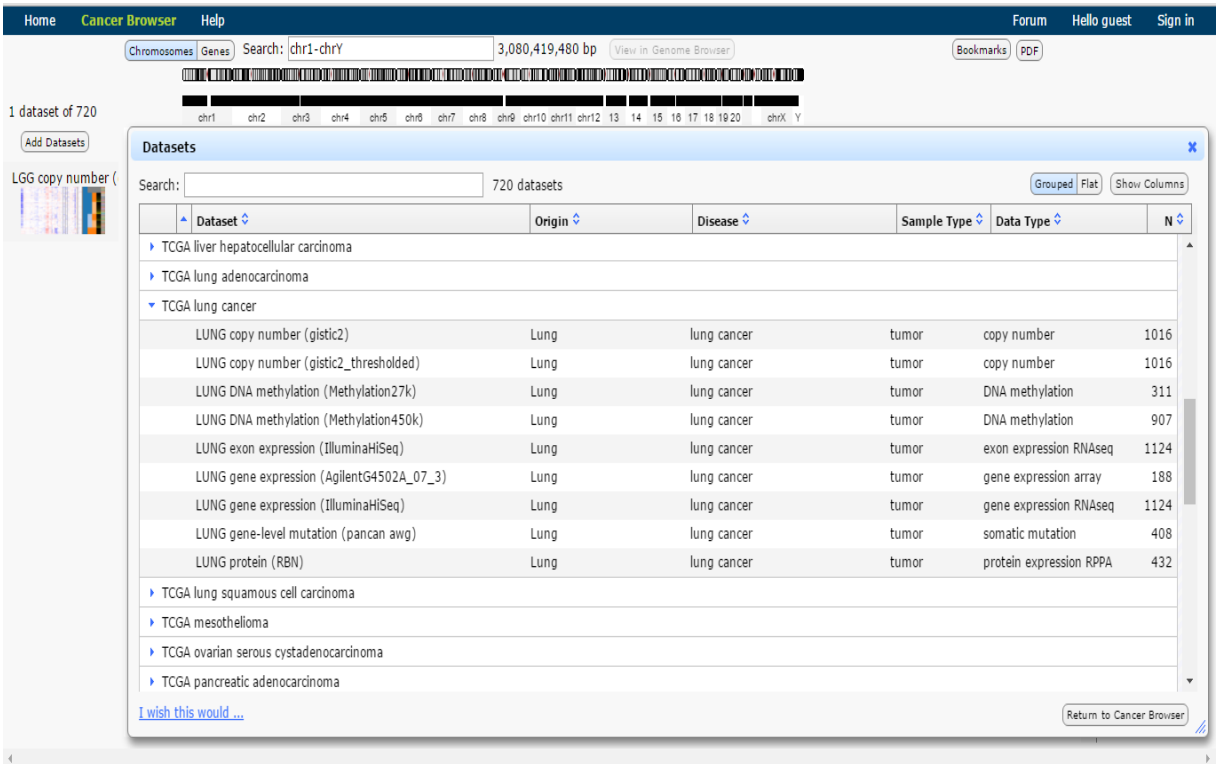



**Fig-15: TCGA** Data portal

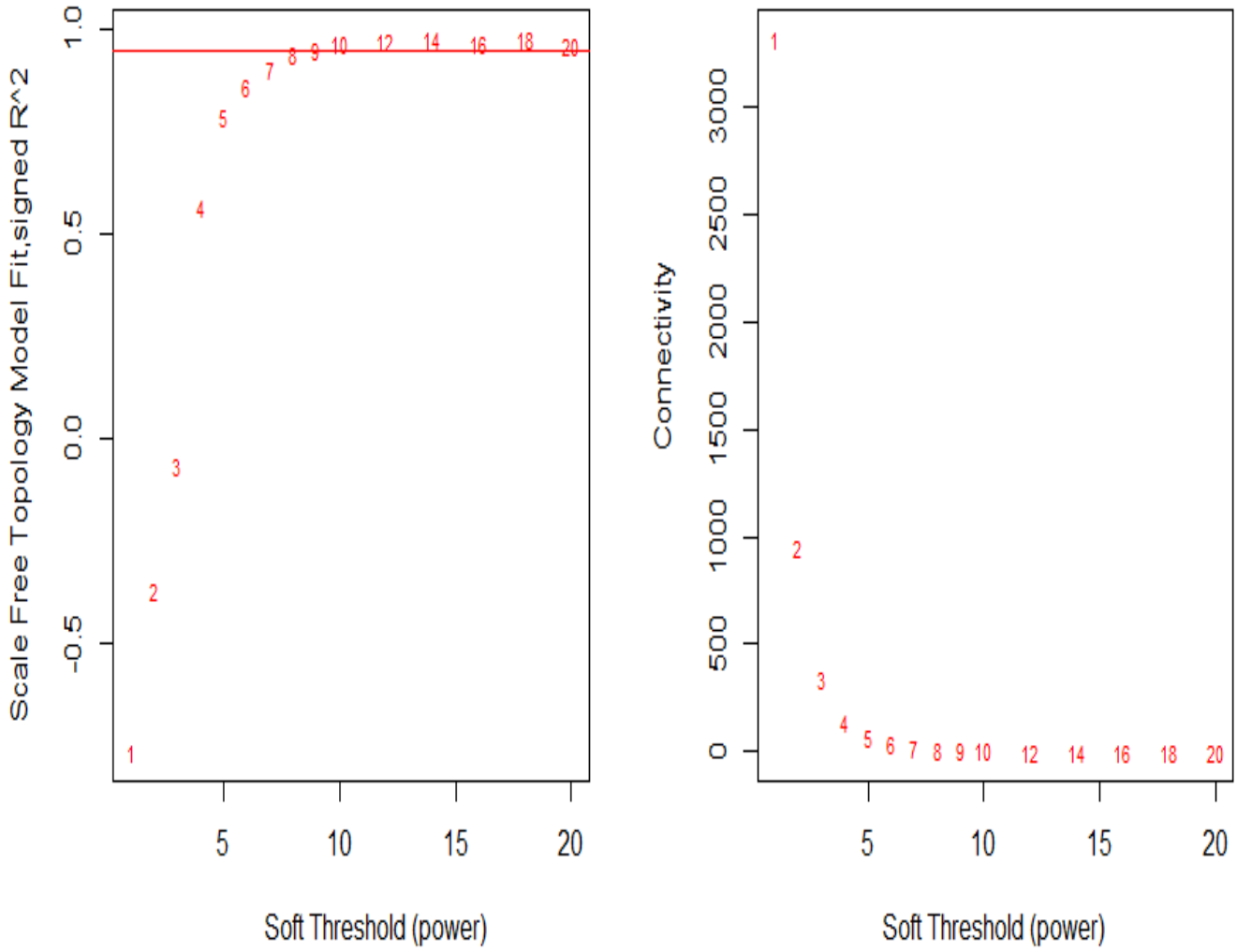



**Fig :** Soft thresholding

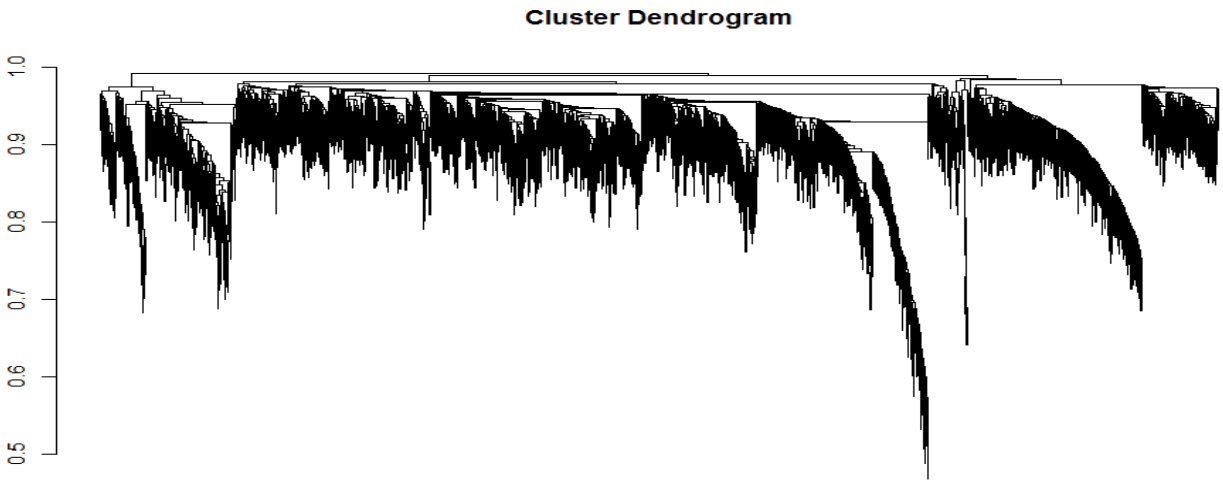



**Fig:** Cluster Dendogram

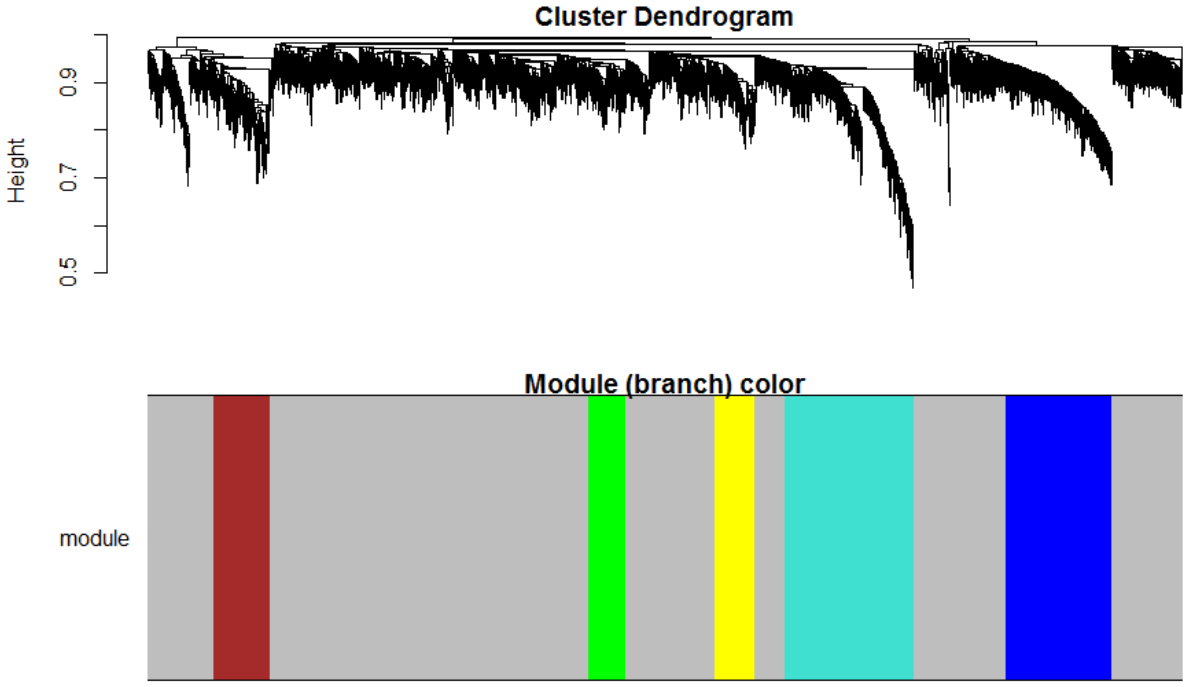


**Fig :** Hierarchical Clustering and Module finding in Microarray dataset

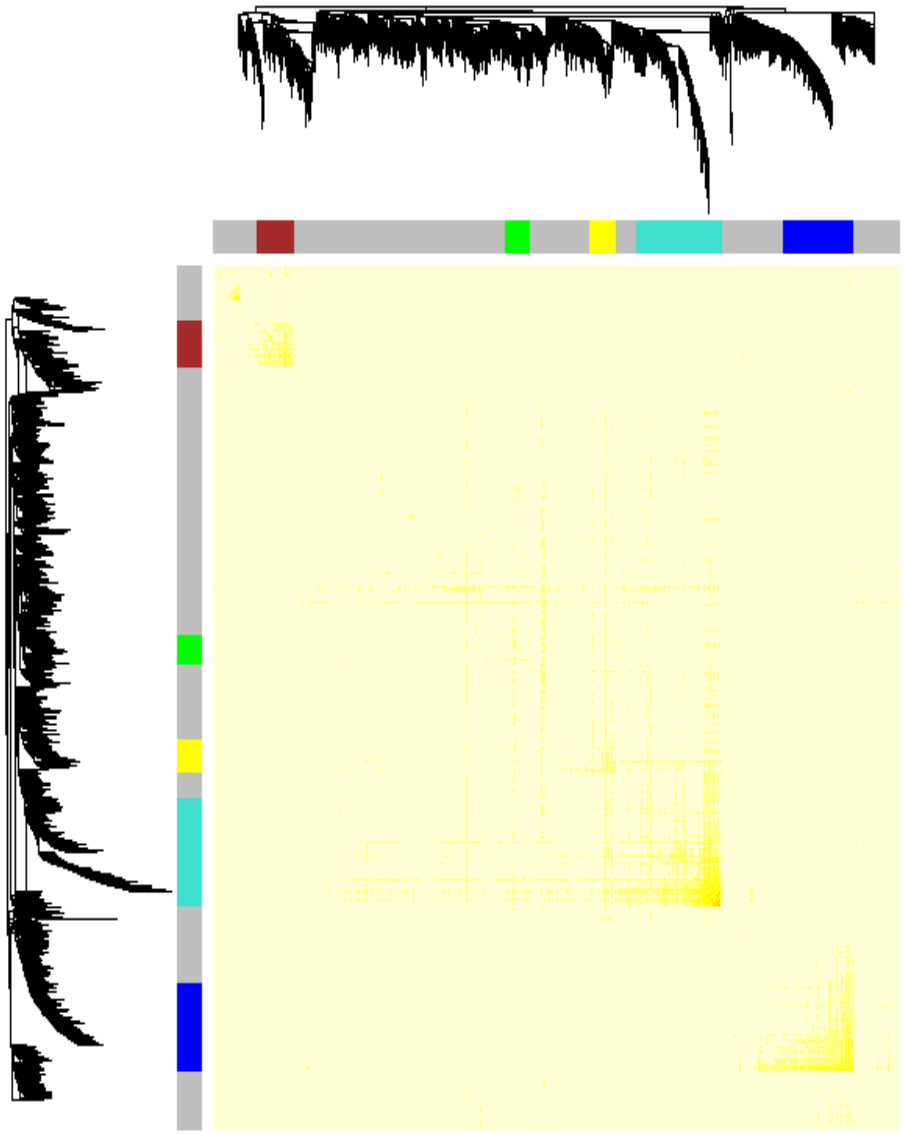


**Fig:** Network heatmap plot

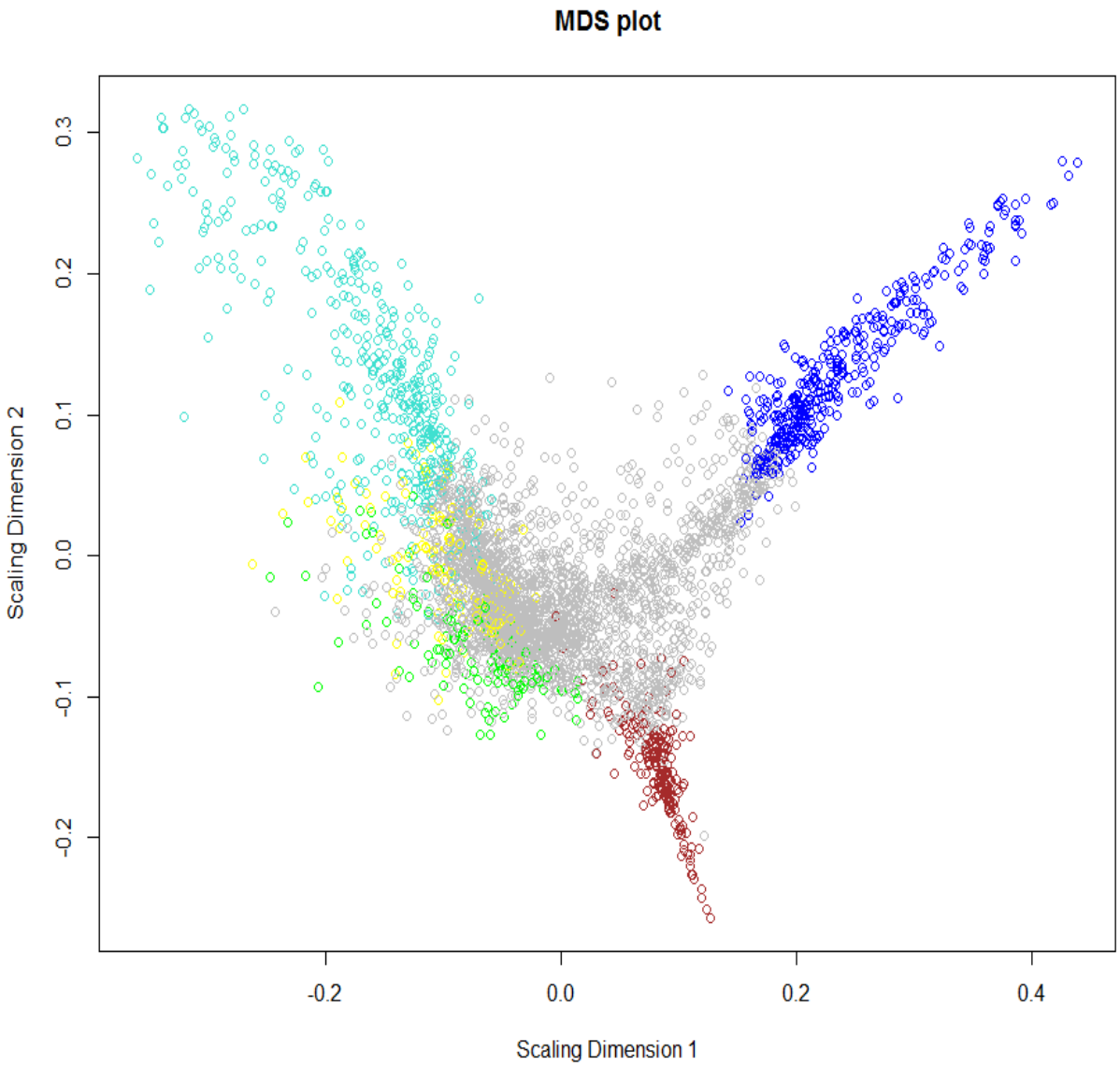



**Fig : Classical MDS**

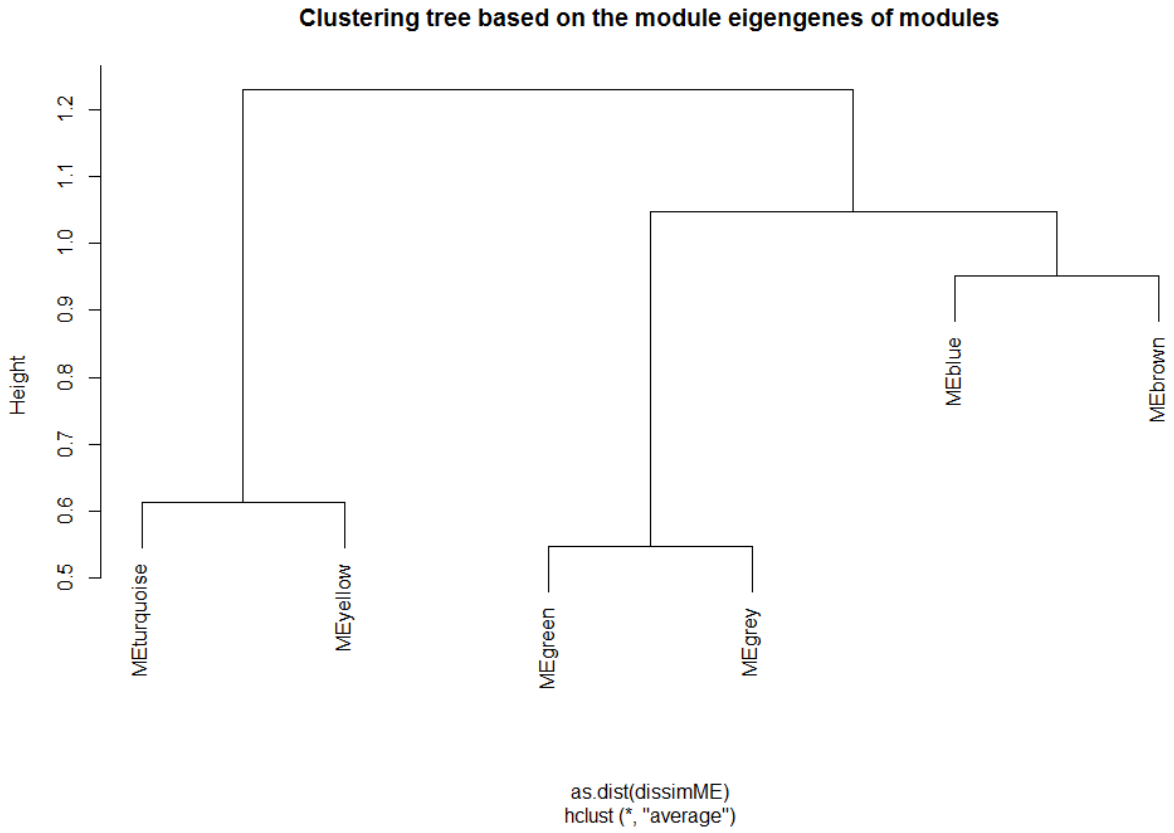



**Fig :** Module eigengene dendrogram

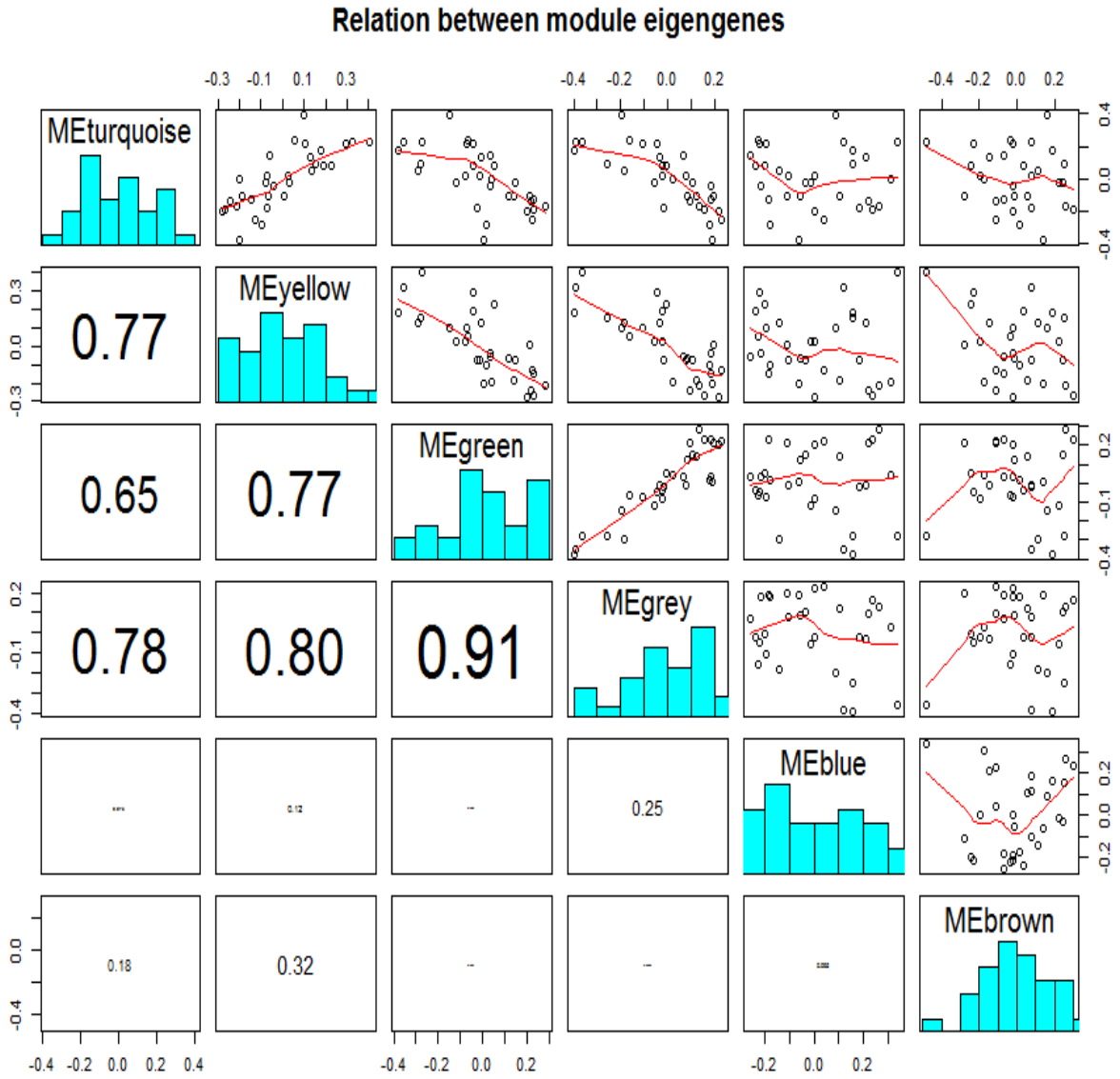



**Fig :** Relation between Eigengenes module

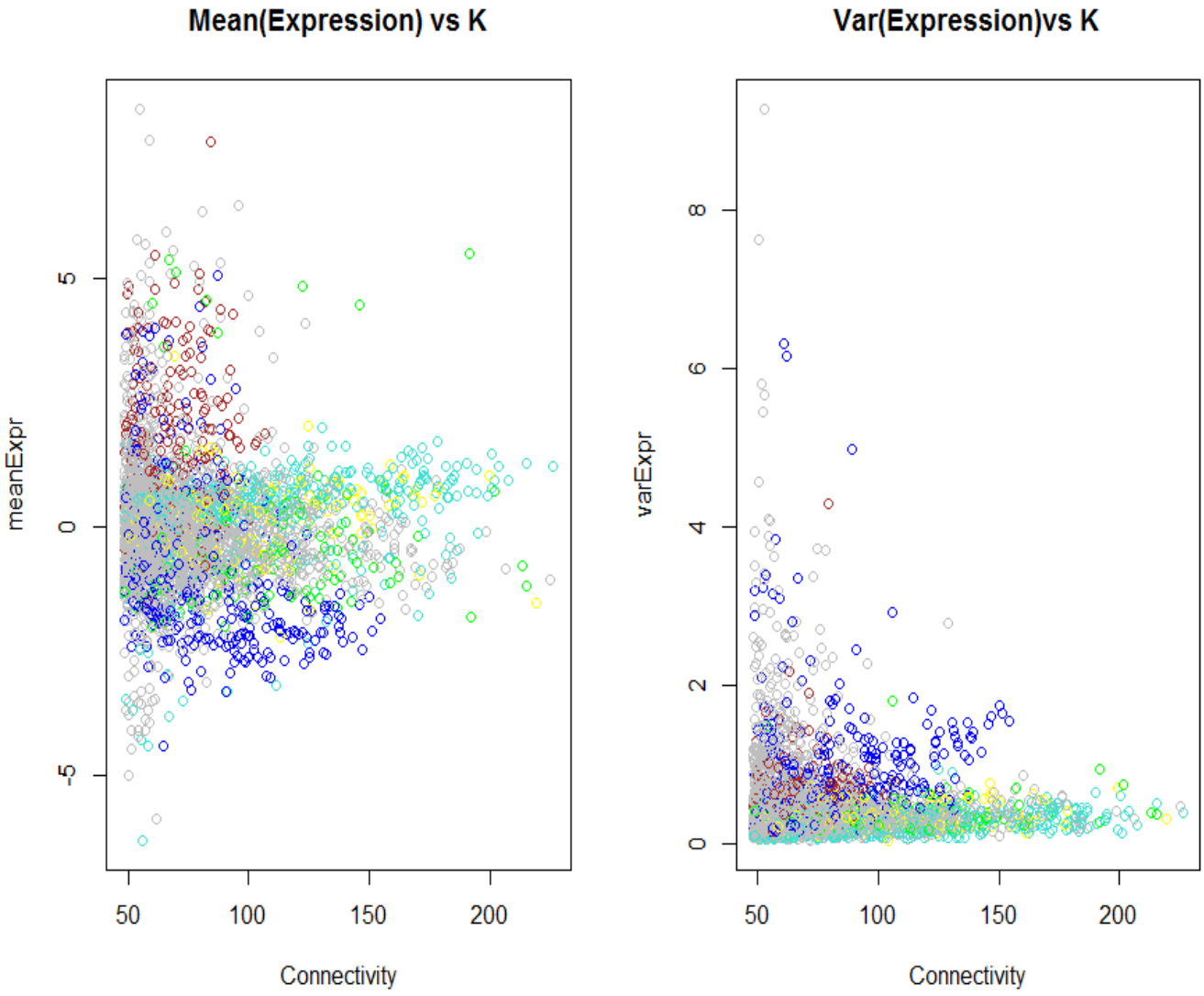



**Fig** : a) Connectivity related to mean gene expression  b) Connectivity related to variance gene expression

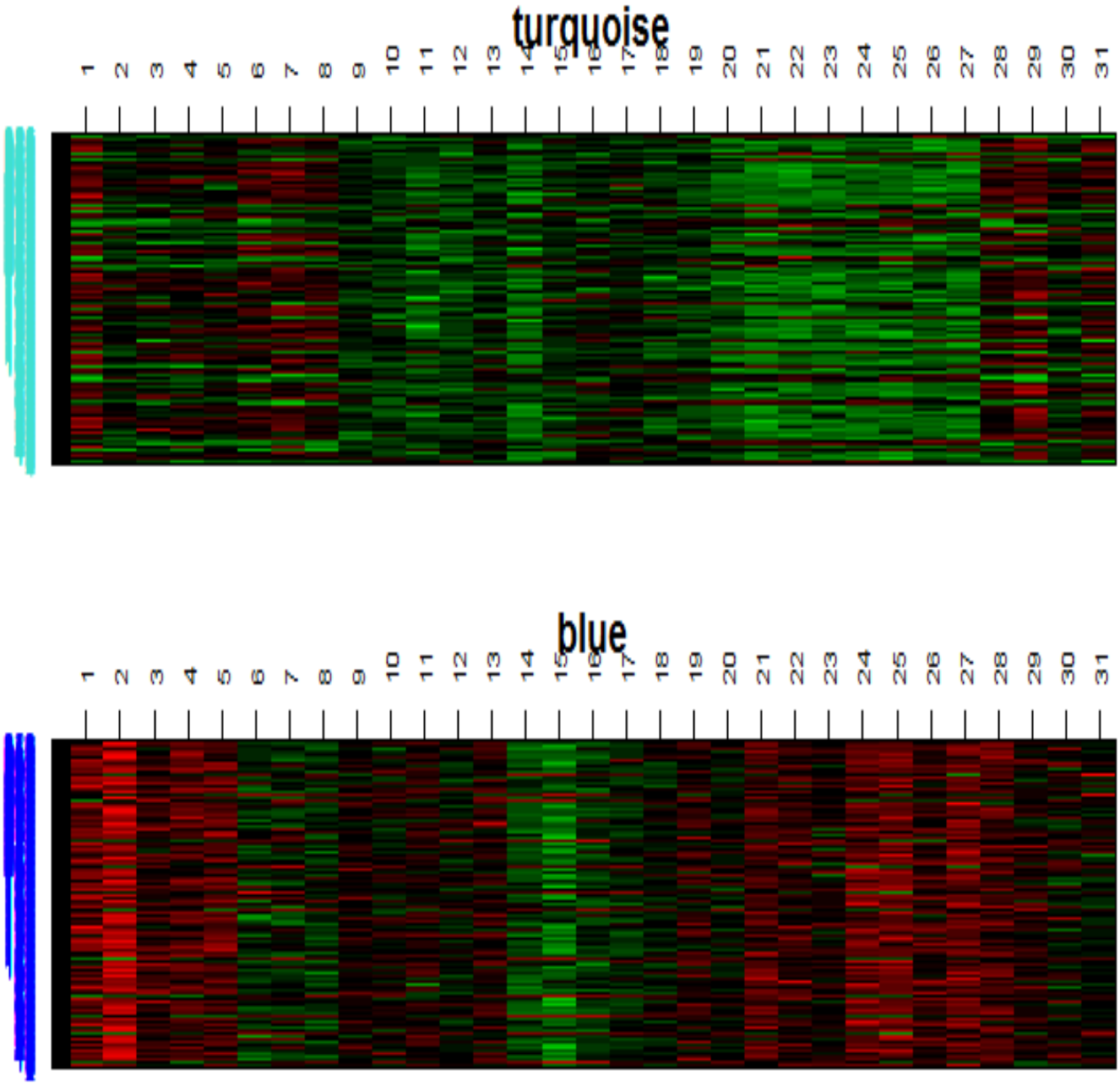



**Fig:** Modules results in characteristic band structures since the corresponding genes are highly correlated.

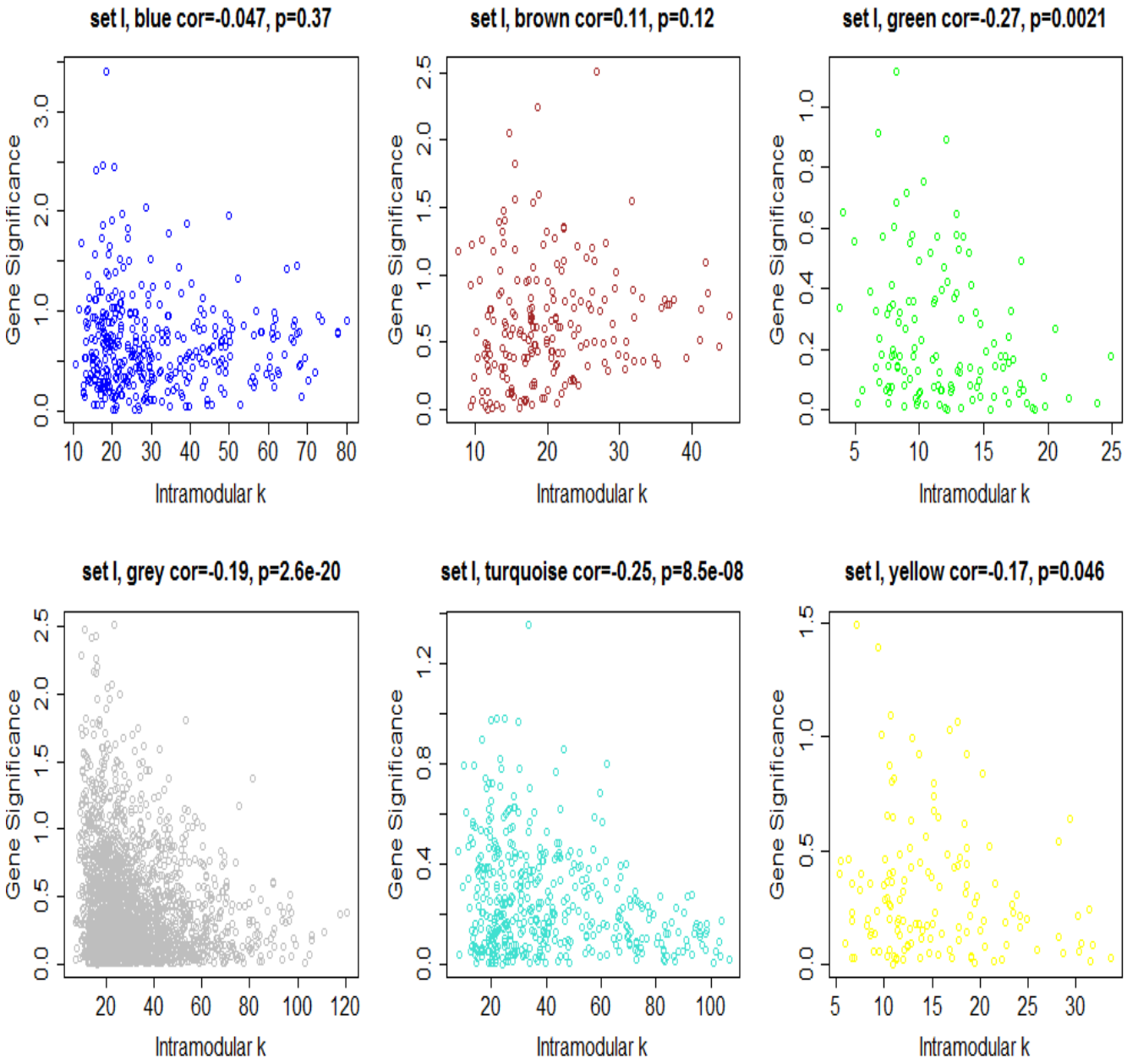



Fig: Gene significance vs intramodular connectivity

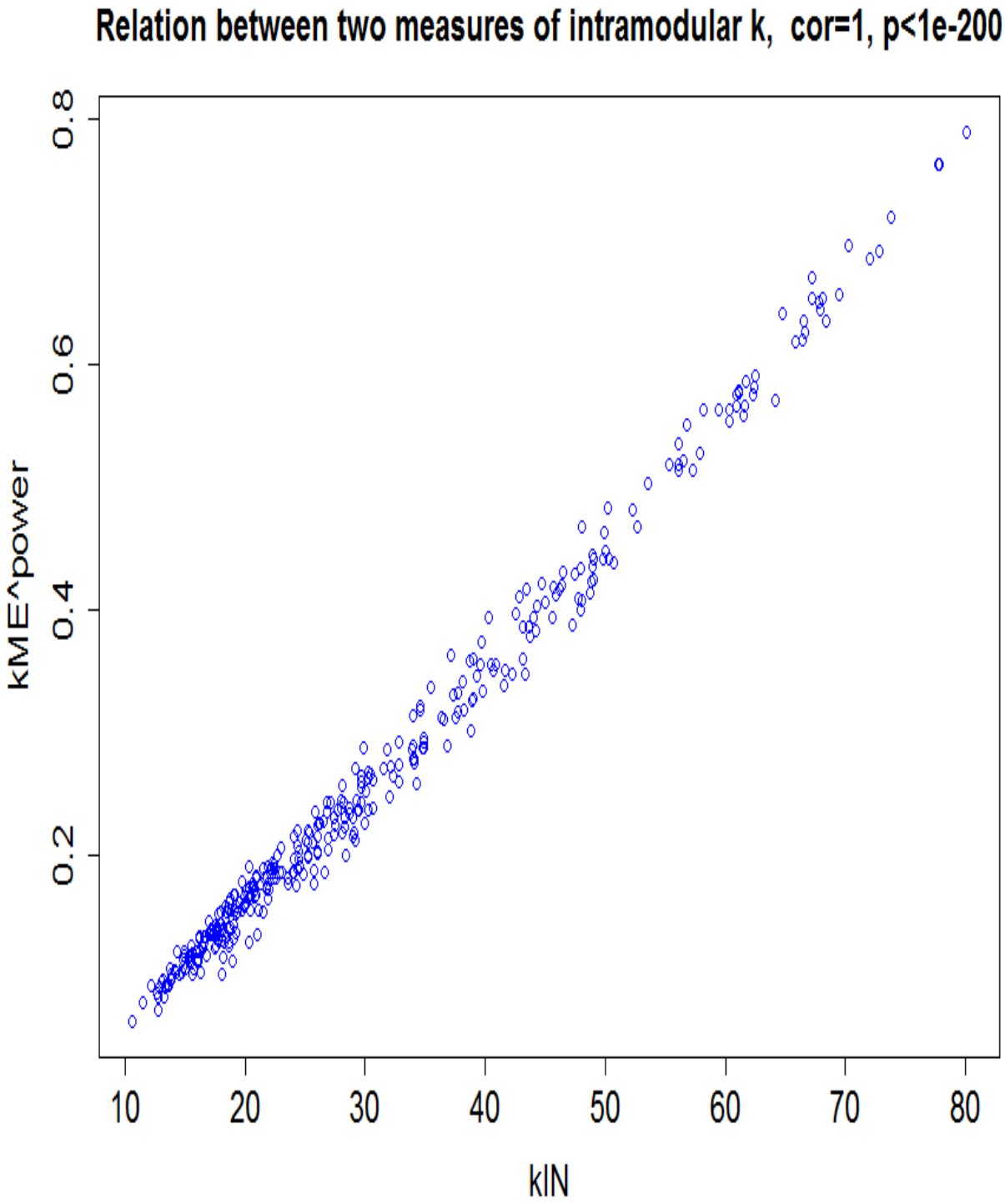



**Fig:** Relation between two measures of intramodular k

### 3.4.3 Result of Simulation:

Among all the modules, only 2 genes were found that satisfied both conditions with high intramodular connectivity as well as a cox p-value of less than 0.05. These are TTk, C6orf173, CENPE and DCC1

The TTk gene has shown to be an oncogene in human cancers[6]. TTk is shown to be overexpressed at the mRNA and protein levels in established NSCLC cell lines as well as primary NSCLC tumors. Also, TTk overexpression was found in 83% of primary NSCLC tumors. This biological finding was vindicated by our analysis which showed TTk gene to not only be highly connected but also significant in terms of patient survival. Moreover, TTk gene overexpression has shown to be associated with poor prognosis in patients with NSCLC[8].

This gene encodes a dual specificity protein kinase with the ability to phosphorylate tyrosine, serine and threonine. Associated with cell proliferation, this protein is essential for chromosome alignment at the centromere during mitosis and is required for centrosome duplication. It has been found to be a critical mitotic checkpoint protein for accurate segregation of chromosomes during mitosis. Tumorigenesis may occur when this protein fails to degrade and produces excess centrosomes resulting in aberrant mitotic spindles. Alternative splicing results in multiple transcript variants.

Again, NSCLC patients whose tumors exhibit strong C6orf173 staining had a poorer prognosis than patients whose tumors showed weak. C6orf173 plays a central role in assembly of kinetochore proteins, mitotic progression and chromosome segregation. It is one of the inner kinetochore proteins, with most further proteins binding downstream. Required for normal chromosome organization and normal progress .

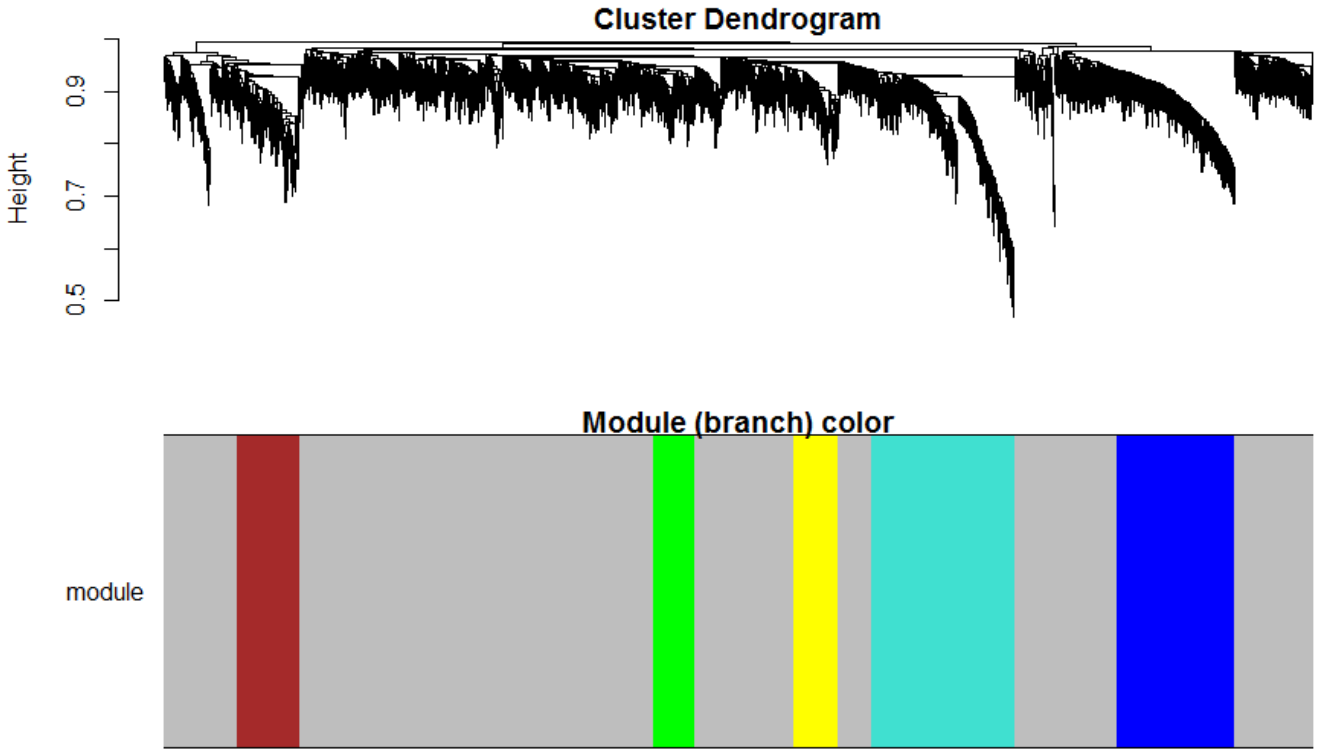
Centromere-associated protein E (CENPE) have entered Phase I and II clinical trials either as monotherapies. As molecularly targeted therapy continues to advance in lung cancer, racial differences in specific genetic/genomic alterations can have an important impact in the choices of therapeutics and in our understanding of the drug sensitivity/resistance profile. Commonly identified genetic/genomic alterations such as missense or nonsense mutations, small insertions or deletions, alternative splicing, and chromosomal fusion rearrangements were discussed on CENPE.

No DCC1 staining which also supports our finding of DCC1 being prognostically significant.
The DCC1 gene has also shown to be a useful biomarker for aggressive breast tumors[9]. It was seen that siRNA-mediated knockdown of DCC1 significantly reduced cell proliferation in ER-

negative normal breast and cancer lines. This implied DCC1 is essential for both normal and cancer cell survival.Moreover, overexpression of DCC1 was also correlated with poor survival. *DCC* candidate tumor suppressor gene has been mapped in this region, mutation and expression of the *DCC* gene were examined in 46 lung cancer cell lines, consisting of 14 small cell lung carcinomas (SCLCs) and 32 non-small cell lung carcinomas (NSCLCs), to elucidate the pathogenetic significance of *DCC* alterations in human lung carcinogenesis. A heterozygous missense mutation was detected in a NSCLC cell line, Ma26, while homozygous deletion was not detected in any of the cell lines. We did a gene enrichment analysis on the genes of the Green and Blue module as these were the top 2 significant modules for our purpose. We used David Bioinformatics Resources 6.7 web interface[16] tool to do the gene enrichment analysis.

### 3.4.4 Comparisons between two dataset:

Now, we wanted to verify that we can reproduce these modules and their emergence was not a one time thing. Hence, we repeated the whole analysis on another microarray dataset with the same set of genes. We get 5 modules from our microarray dataset (given below). To see how these modules were connected, we can create a clustering tree of the eigengenes for each module



### 3.4.5 Result Verification:

Many significant genes that were found in the Microarray data set were found again in the RNA-seq data set. For each module here, we have circled it with the color of the RNA-seq module where its genes can be found. These results increase our confidence in the original findings and their biological interpretations.

# Chapter 4

# Conclusion

## 4. Summary of the research

At the beginning , our goal was to do a thorough exploratory analysis of lung squamous carcinoma and try to find genes and modules that are significant in terms of survival time as well as playing a crucial role in cancer growth. Our results gave us two genes which are highly connected as well as prognostically significant. The literature available on these two genes so far support our findings. Moreover, the highly connected genes in the significant modules are also shown to be involved in processes that are crucially connected to cancer cell growth. At the same time, the gene enrichment analysis showed some functions like involvement with glycoprotein or phosphoprotein where the immediate biological connection with lung cancer is not really clear. Biological experiments targeting these findings in the future might confirm or negate our findings.

# 5. References


1. Langfelder, P. & Horvath, S. (2008) *WGCNA: an R package for weighted correlation network analysis.*
2. Horvath, S., et al. "Analysis of oncogenic signaling networks in glioblastoma identifies ASPM as a molecular target." *Proceedings of the National Academy of Sciences* 103.46 (2006): 17402-17407.
3. A Research for Weighted Gene Co-expression Network Model.Jun Wang, Weiping Wang, Wen Liu, Zhong Zhou, Xiaoying Wang.
4. Yip A, Horvath S: Gene network interconnectedness and the generalized topological overlap measure. BMC Bioinformatics 2007, 8:22..
5. Li A, Horvath S: Network Neighborhood Analysis With theMulti-node Topological Overlap Measure. Bioinformatics 2007,23(2):222-231
6. Langfelder P, Zhang B, Horvath S: Defining clusters from a hierarchical cluster tree: the Dynamic Tree Cut package for R.Bioinformatics 2008, 24(5):719-720.
7. Ghazalpour A, Doss S, Zhang B, Plaisier C, Wang S, Schadt E, Thomas A, Drake T, Lusis A, Horvath S: Integrating Genetics and Network Analysis to Characterize Genes Related to Mouse Weight. PloS Genetics 2006, 2(8):e130
8. Fuller T, Ghazalpour A, Aten J, Drake T, Lusis A, Horvath S:Weighted Gene Co-expression Network Analysis StrategiesApplied to Mouse Weight. Mammalian Genome 2007,6(18):463-472
9. Oldham M, Horvath S, Geschwind D: Conservation and Evolution of Gene Co-expression Networks in Human and Chimpanzee Brains. Proc Natl Acad Sci USA 2006, 103(47):17973-17978.
10. Horvath S, Dong J: Geometric interpretation of Gene Coexpression Network Analysis. PLoS Computational Biology 2008.

11. Dudoit S, Yang Y, Callow M, Speed T: Statistical methods for identifying differentially expressed genes in replicated cDNA microarray experiments. Statistica Sinica 2002.
12. (2006) FDT: fields: Tools for Spatial Data. Tech. rep., National Center for Atmospheric Research, Boulder, CO 2007 [http://www.image.ucar.edu/GSP/Software/Fields]
13. Hu Z, Snitkin ES, DeLisi C: VisANT: an integrative framework for networks in systems biology. Brief Bioinform 2008,9(4):317-325.
14. Shannon P, Markiel A, Ozier O, Baliga NS, Wang JT, Ramage D, Amin N, Schwikowski B, Ideker T: Cytoscape: A Software Environment for Integrated Models of Biomolecular Interaction Networks. Genome Research 2003, 13(11):2498-2504.

15. Mutation and Expression of the DCC Gene in Human Lung Cancer. Takashi Kohno, Takako Sato, Satoshi Takakura, Kimiko Takei, Kaoru Inoue, Michiho Nishioka, and Jun Yokota,2000.
16. *DAVID - Gene enrichment analysis tool.* https://david.ncifcrf.gov/home.jsp