DATA ANALYSIS AND CREATION OF EPIGENETICS DATABASE

Akshay A Desai

Submitted to the faculty of the Bioinformatics Graduate Program in partial fulfillment of the requirements for the degree Master of Science in Bioinformatics in the School of Informatics Indiana University December 2013

Accepted by the Graduate faculty, Indiana University, in partial fulfillment of the requirements for the degree of Master of Science in Bioinformatics

Master's Thesis Committee

_____

Mathew Palakal, Ph.D.

_____

Huanmei Wu, Ph.D.

_____

Xiaowen Liu, Ph.D.

Copyright page

To my parents, Mr. Ashok B. Desai and Mrs. Vidya A. Desai and my family. They raised me, supported me, taught me and loved me. This thesis is dedicated to them.

Acknowledgement

This thesis was made possible due to the guidance and encouragement from many people. So it gives me a great pleasure to thank these people and acknowledge their contribution. I owe a sincere appreciation to my thesis advisor Dr Mathew Palakal for his motivation and immense knowledge. Without his thoughtful guidance, energy and constructive criticism this thesis would have never been possible. I would like to thank Dr Meeta Pradhan for her role in my thesis completion. Besides my lab members, I would also like to thank the other members of my committee Dr Huanmei Wu and Dr Xiaowen Liu for supporting my work, reading my thesis and providing helpful suggestions.

I wish to express sincere appreciation to School of Informatics at Indiana University Purdue University Indianapolis for providing me an opportunity to pursue a bright carrier in bioinformatics.

I thank my family for the constant support, love and blessing. Their teachings have made me the person, I am today.

Akshay A Desai

DATA ANALYSIS AND CREATION OF EPIGENTICS DATABASE

Abstract

This thesis is aimed at creating a pipeline for analyzing DNA methylation epigenetics data and creating a data model structured well enough to store the analysis results of the pipeline. In addition to storing the results, the model is also designed to hold information which will help researchers to decipher a meaningful epigenetics sense from the results made available. Current major epigenetics resources such as PubMeth, MethyCancer, MethDB and NCBI's Epigenomics database fail to provide holistic view of epigenetics. They provide datasets produced from different analysis techniques which raises an important issue of data integration. The resources also fail to include numerous factors defining the epigenetic nature of a gene. Some of the resources are also struggling to keep the data stored in their databases up-to-date. This has diminished their validity and coverage of epigenetics data. In this thesis we have tackled a major branch of epigenetics: DNA methylation. As a case study to prove the effectiveness of our pipeline, we have used stage-wise DNA methylation and expression raw data for Lung adenocarcinoma (LUAD) from TCGA data repository. The pipeline helped us to identify progressive methylation patterns across different stages of LUAD. It also identified some key targets which have a potential for being a drug target. Along with the results from methylation data analysis pipeline we combined data from various online data reserves such as KEGG database, GO database, UCSC database and BioGRID database which helped us to overcome the shortcomings of existing data collections and present a resource as complete solution for studying DNA methylation epigenetics data.

Table of Contents

Lists of tables

List of Figures

Chapter 1 Introduction

1.1 What is Epigenetics?

The genotypic make up of monozygous twins is identical. However, a study[1] on monozygous twins proved that they are not identical as they do not show same disease susceptibility. The study also observed that this deviation in susceptibility increases with the ageing of twins. Now the cause of most diseases is attributed to the alteration in the expression level of genes. This alteration affects the normal functionality of the protein machinery involved in the concerned affected biological process, hence the disease. In monozygotic twins if the genetic makeup is not altered than what are the possible reasons for the variation in disease susceptibility? What are the factors responsible for affecting the gene expression levels other than mutations in DNA sequences? A substantial part of these questions is answered by a new field called “Epigenetics”.

Epigenetics was first defined by Conrad Waddington in 1940. According to Conrad Waddington epigenetics is “the interactions of genes with their environment which bring the phenotype into being”. As biological terms have different meanings for different people[2], epigenetics is no exception to this tradition. In contrast to Conrad’s definition, Arthur Riggs et al defined epigenetics as “the study of mitotically and/or meiotically heritable changes in gene function that cannot be explained by changes in DNA sequence”[3]. The two definitions mentioned explain important aspects of epigenetics but do not cover the field in totality. One definition explains epigenetics from a developmental biologist perceptive and the other one defines what epigenetics is not i.e. an inheritance of mutations. But the definition from an article[2] in nature very well summarizes the concept. It states that epigenetics is “the structural adaptation of chromosomal regions so as to register, signal or perpetuate altered activity states”.

Epigenetics marks determine the functional output of the information that is stored in the genome. These marks are the link between prenatal and postnatal exposures and the phenotypic changes happening later in life. This connection is beautifully explained by a BBC T.V science programme. It states that “at heart of this new field is a simple but contentious idea- that genes have ‘memory’. That the lives of your grandparents - the air they breathed, the food they ate, even the things they saw - can directly affect you, decades later, despite your never experiencing these things yourself. And that what you do in your lifetime could in turn affect your grandchildren.” Epigenetics tries to explain genes beyond the DNA. The word epigenetics also literally means ‘above the genetics’ which is now often used to explain gene expression changes that occur without the change in DNA sequence. It explains the heritable changes in gene expression or cellular phenotype caused by the factors other than changes in DNA sequences. These factors could be nutrition, stress or other environmental influences which causes a gene to turn on or off. These changes may remain through cell divisions for the remainder of the cell’s life and may also last for generations. However, there is no change in the underlying DNA sequence of the organism; instead, non-genetic factors cause the organism’s genes to behave differently.

DNA methylation and histone modification are the two important molecular mechanisms playing a major role in epigenetic regulation of genes[4]. These modifications affect both DNA and chromatin.

1.2 Important parts of epigenetic machinery.

Most widely studied phenomenon related to epigenetics is DNA methylation. DNA methylation occurs at the carbon-5 position of cytosine in CpG dinucleotides. Histone modifications is the next most common phenomenon studied related to epigenetics. Histone modifications are the post-translational modifications which cause

the changes to the chromatin packaging of DNA. Apart from these two commonly studied mechanisms other epigenetic controls include regulation by microRNAs which are non-coding RNAs and processes that manage higher-level packaging of chromatin within the nucleus.

There are regions in human genome with more than 200 bases which consist of at least 50% G+C content and at least 0.6 ratio of observed to statistically expected CpG frequencies. These CpG dinucleotides clusters are called CpG islands. The “p” in CpG stands for “phosphodiester bond”. This bond is present between cytosine and guanine which are the “C” and “G” in CpG. Cytosine methylation is the most extensively studied phenomenon in context with epigenetic modifications. Modifications in the methylation state of CpG island are epigenetically important because about 60% of human gene promoters are associated with CpG islands. These promoter regions are usually un-methylated in normal cells but nearly 6% of them become methylated. This methylation occurs in a tissue-specific manner during early development or in different tissues[5]. CpG-island methylation is commonly associated with silencing of genes which causes the reduction in the expression level of these genes. This kind of methylation is the only epigenetic modification that directly affects the DNA, hence the name “DNA methylation”. DNA methylation causes the replacement of hydrogen atom of the cytosine base by a methyl group which affects the accessibility of transcription binding sites thus causing the silencing of genes. In cancer, epigenetic modification due to DNA methylation causes the silencing of tumor suppressing genes.

Histone modification is another important epigenetic factor. Nucleosome at the core contains a histone octamer comprising of 2 copies each of H2A, H2B, H3 and H4. A DNA of 147-bp segment is wrapped around the histone octamer in 1.65 turns. Neighboring nucleosomes are separated by a ~50 bp of DNA. H1 is the histone which binds to the linker DNA, binding off the nucleosome at the location where DNA enters

and leaves, hence it is called the linker histone. The core histones are modified at their amino-terminal tails by acetylation, phosphorylation, methylation and ubiquitylation. These modifications play an important role in gene regulation by defining the gene activity based on the combined acetylation, phosphorylation and methylation status and occur mostly post-transcriptionally.

There are 3 classes of chromatin-remodeling proteins in mammalian cells. These complexes are SWI/SNF/Brm, ISWI and Mi-2/NuRD and contain different catalytic ATPase subunits with associated proteins. The enzymes involved in chromatin-remodeling use the ATP hydrolysis energy to influence the structure of nucleosome and the DNA wrapped on it. Such influence of these proteins affects the accessibility of chromatin to various chromatin proteins that control transcription, DNA replication, recombination and other biological processes.

1.3 Epigenetics: Hope for Cancer cure.

Global changes in DNA methylation, histone modification and chromatin-modifying enzyme expression profiles characterizes the cancer epigenome[6]. Cancer cells acquire a specific hypomethylation and hypermethylation patterns at the CpG islands of the promoter regions of genes. Hypomethylation at promoter regions of oncogenes activates their aberrant expression while hypermethylation at promoter regions leads to inactivation of tumor suppressor genes. Chromosomal instability, translocations and gene disruption is caused by the global hypomethylation at repetitive sequences[7]. Hypomethylation in the promoter regions of genes like S100P in pancreatic cancer, SNCG in breast and ovarian cancers and MAGE and DPP6 in melanomas is a well-studied phenomenon. Hypermethylation on the other hand is mainly observed at specific CpG islands and affects genes involved in the main cellular

pathways like DNA repair, vitamin response, Ras signaling, cell cycle control, p53 network and apoptosis.

In cancer cells a global reduction of mono-acetylated H4K16 is observed as an important histone modification. This loss in acetylation is mainly due to the overexpression of entities belonging to Sirtuin family of proteins[8]. Apart from the global loss of H4K16, there is a variation in the distribution of the histone methyl marks in cancer cells that is observed. This variation is mainly caused due to the alteration in the expression of histone methyltransferases and demethylases. Histone phosphorylation, too, according to current studies is found to be relevant for cancer studies. JAK2, a non-receptor tyrosine kinase, is involved in hematological malignancies by getting activated through chromosomal translocations and point mutations. In addition to methylation and histone modifications, all families of chromatin remodelers are associated with cancer. Promoter hypermethylation of MLH1 in colon cancer is observed which proves the involvement of nucleosome remodeling in the transcriptional down regulation by promoter hypermethylation.

Epigenetic therapy is targeted towards reprogramming the network of chemical changes that alter the functioning of cancer cells DNA rather than destroying them by disrupting their DNA or affecting important cancer pathways. The possibility for therapy has arisen due to the reversible nature of epigenetic changes. Hence the main aim of therapy is to restore ‘normal epigenome’. The first epigenetics drugs to be proposed for cancer treatment are DNA methylation inhibitors. Drugs such as 5-Aza-CR (azacitidine) and 5-aza-CdR (decitabine) which are DNA methylation inhibitors are approved by FDA for treating myelodysplastic syndromes[9]. A recent clinical trial conducted by researchers at John Hopkins institute proved the effectiveness of low-dose azacitidine combined with entinostat which is another epigenetic drug, for treating patients with advanced lung cancer. Apart from lung cancer the team also proved the effectiveness of

low doses of these drugs with antitumor effects in cell lines and in mouse models for different cancers such as leukemia, breast and colon cancer.

Often loss of histone acetylation is also attributed to the aberrant gene silencing in cancer. HDAC inhibitors have been proved to demonstrate anti-tumorigenic effects which include activities such as growth arrest, apoptosis and the induction of differentiation. These inhibitors help in re-establishing the normal histone acetylation patterns and reactivating silenced tumor suppressor genes[10]. Suberoylanilide hydroxamic acid (SAHA) is one such HDAC inhibitor which has been proved for its use in clinic for treating T cell cutaneous lymphoma. Combinatorial cancer treatment strategies have also been explored which includes use of both DNA methylation and HDAC inhibitors for treating cancer. The study exploring the combinatorial effects showed that the activation of certain tumor suppressor genes was seen only when 5-Aza-CdR was administered in combination with trichostatin A. A combination of 5-Aza-CdR also enhanced the anti-tumorigenic activity of depsipeptide on leukemic cells. Similar combinatorial effect was observed when phenyl-butyrate and 5-Aza-CdR where used together to demonstrate the reduction of lung tumor formation in mice. Apart from DNA methylation and HDAC inhibitors, HMT inhibitors and miRNAs are also explored for epigenetic therapy and hold a great potential. Thus epigenetics has defined a new way and inculcated a hope in the war against cancer. It has provided a fresh approach in understanding the underlying principle for cancer treatment.

1.4 Network Biology

A systematic catalogue of all molecules and their key interactions in a desired cellular system is important for any post-genomic biomedical research. Need to understand the interactions between these molecules and the functionality of these interactions lead to the development of “network biology” field. The advances in network

biology have helped to build new conceptual framework that has revolutionized our understanding about the biology and disease pathologies. It has been well established now that a complex interaction between different cellular components such as proteins, DNA, RNA and small molecules define a biological characteristic. Thus it is important to understand the dynamics and structure of complex intercellular interactions that are the underlying principles in formation of structure and function of a living cell.

Current high-throughput data analysis techniques for understanding epigenetic processes generate different types of interactions which include types such as protein-protein interaction, metabolic, signaling and transcription-regulatory networks. Each molecule is represented as a node in the network while the interactions between different molecules are represented by the edges of a network. A network of these networks helps us to understand the behavior of cell. Network biology quantifies different aspects of a given biological network which helps to characterize various biological systems. Following are some of the important network measures that let us compare and characterize different complex networks:

a. Degree: The degree of a node in a network is defined as the number of connections or edges the node has to other nodes. In a directed network every node has two types of degrees; in-degree and out-degree. In-degree defines the number of incoming edges while the out-degree defines the number of outgoing edges.

b. Degree distribution: It is the fraction of nodes in the network with degree k. Thus, the degree distribution P (k) is calculated by obtaining the nodes with degree k and dividing by the total number of nodes. This kind of distribution allows differentiation between different classes of the network.

c. Shortest path and mean path length: Shortest path is the path with the smallest number of links between any selected nodes which are the nodes of interest. Whereas

mean path length is the average of the shortest paths between all pairs of nodes and provides a measurement which facilitates the overall network navigability.

d. Clustering coefficient: It is defined as the measure which helps to calculate the degree to which nodes in a graph tend to cluster together.

Network biology has become an integral part of any biological data analysis. It has helped to understand various complex biological machineries and produce some important inference mechanisms. The field has great potential to interpret epigenetic data which is been generated now-a-days using different high-throughput processes.

Chapter 2 Background

2.1 Current approaches for analyzing epigenetic data.

A genome wide mapping of epigenetic information is done by using techniques which use a three-stage design. In first stage epigenetic information is obtained by biochemically converting it into genetic information. This conversion is achieved by enriching specific genomic regions which show aberrant epigenetic modifications. Second stage involves employment of high-throughput data generation techniques such as microarray and sequencing techniques. These techniques generate data which help to interpret epigenetic modifications and make a scientific prediction or analysis of the process. The third stage includes computational algorithms to make actual inference from the data generated from microarray and sequencing methods.

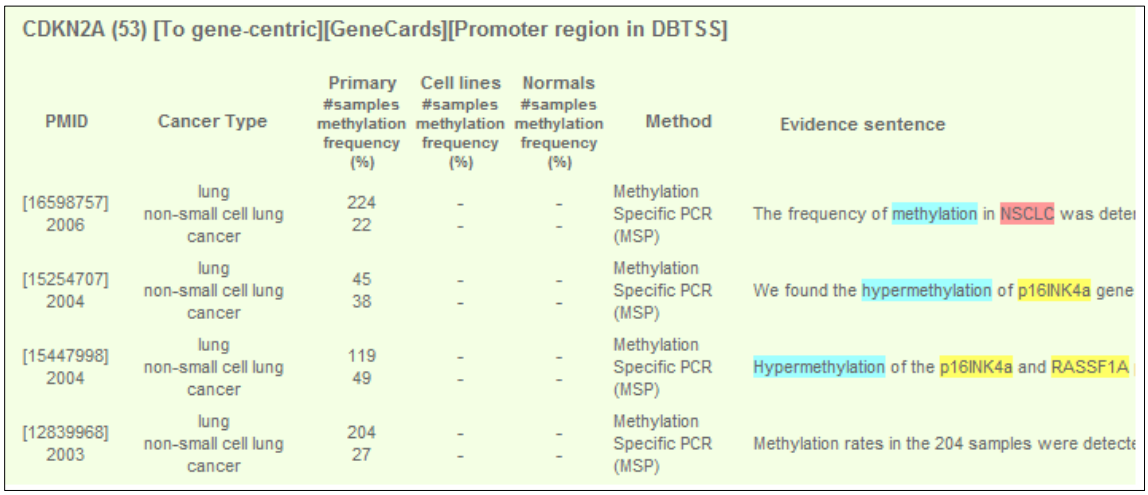
Bisulphite treatment of DNA is the popular method for detecting DNA methylation. It reproducibly alters un-methylated cytosines to uracil, leaving methylated cytosines unchanged. This resolution of treatment yields single-nucleotide resolution information about the methylation status of a sequence of DNA. In addition combining this methodology with sequencing technologies and amplifying it with methylation-specific PCR allows investigating DNA methylation even at low quantities. Techniques such as ChIP-on-chip which are based on chromatin immunoprecipitation have recently made a major contribution in epigenomics profiling of cancer cells. Identification of histone modification is more challenging compared to DNA methylation assay[6]. The standard technique for finding histone modification is mass spectrometry but the technique is complex and difficult to implement genome-wide. ChIP-on-chip with genomic platforms is also used for analyzing histone modification data.

Projects such as AHEAD Task, ENCODE Project, HEP Project Consortium have developed ground breaking techniques which have improved large-scale experimental

methods and have introduced new bioinformatic methods for analyzing epigenetic data generated by the methods. One of the projects ENCODE which was aimed to mapping functional elements in the human genome has greatly contributed in standardizing the epigenome data analysis process. The first step in the process involves unsupervised segmentation of chromatin data which is based on the wavelet smoothing and hidden Markov models[11]. The second step is the joint statistical analysis on the datasets from ENCODE pilot phase. This is an exploratory procedure on a large and heterogeneous datasets which hold a good amount of epigenetic information. Third step included annotation of functional promoters using alternative prediction methods which were developed and evaluated for the project. This step proves the ability of epigenetic data to improve the accuracy of promoter annotation. Fourth the overlap between two sets of genomic features is assessed by determining the significance of overlap by the statistical test developed by the researchers working in ENCODE project. This statistical test helps them to evaluate the significance of overlap between regions such as CpG islands and un-methylated genomic regions. The test also helps to obtain more realistic P-values compared to other randomization-based methods which give an over-estimated significance[11]. Fifth the epigenome datasets from the ENCODE project are incorporated into UCSC Genome Browser. UCSC genome Browser provides visualization and retrieval of these genomic and epigenomics datasets. Thus, these are the steps that were standardized by ENCODE project for analyzing epigenomics data.

In addition to the techniques mentioned, bioinformatics methods like text mining and data mining combined with system's biology also play an important role in analyzing epigenomics data. These techniques are widely used for downstream filtering, processing and annotation of the data.

2.2 Available epigenetic data resources: Features and Limitations



CDKN2A (53) [To gene-centric][GeneCards][Promoter region in DBTSS]

PMID	Cancer Type	Primary #samples methylation frequency (%)	Cell lines #samples methylation frequency (%)	Normals #samples methylation frequency (%)	Method	Evidence sentence
[16598757] 2006	lung non-small cell lung cancer	224 22	- -	- -	Methylation Specific PCR (MSP)	The frequency of methylation in NSCLC was dete
[15254707] 2004	lung non-small cell lung cancer	45 38	- -	- -	Methylation Specific PCR (MSP)	We found the hypermethylation of p16INK4a gene
[15447998] 2004	lung non-small cell lung cancer	119 49	- -	- -	Methylation Specific PCR (MSP)	Hypermethylation of the p16INK4a and RASSF1A
[12839968] 2003	lung non-small cell lung cancer	204 27	- -	- -	Methylation Specific PCR (MSP)	Methylation rates in the 204 samples were detecte

Figure 1: PubMeth results for querying with specific cancer type

PubMeth, MethyCancer, MethDB and NCBI's Epigenomics database are the main data resources that are currently available providing useful information for epigenetic research. PubMeth is a database consisting of genes which are found to be methylated in specific cancer types. It is a cancer methylation database using text-mining from Medline/PubMed abstracts in addition to manual reading and annotation of preselected abstracts for extracting methylated genes[12]. The interface is designed to rank, summarize and present data in such a way that the information present in database is easily accessible. In PubMeth database a query can be either genes based or can be cancer type based. Thus PubMeth tries to collect information on methylated genes in different cancer types from all possible available literature data and provides an interface which summarizes the collected data.

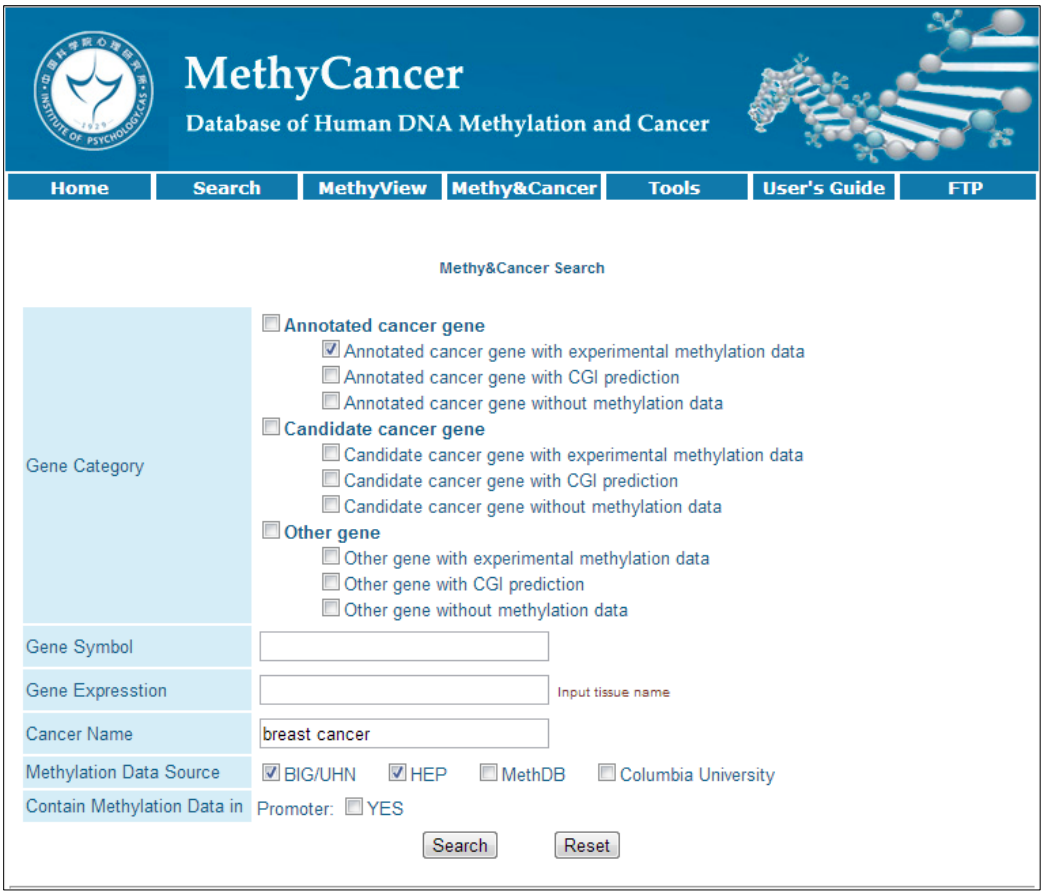


Figure 2 Search panel for MethyCancer database

MethyCancer is a database holding information on the relationship between DNA methylation, gene expression and cancer. This interplay between different facets of epigenetics is made available in the database through integration of large-scale data, its production and mining. The main data sources for the database are public resources which are accompanied by manual curation of the data obtained. In addition to public sources, the Cancer Epigenome Project in China also acts as a major contributor of experimental data to the database. Database holds 4 main types of data: CGI clones and global CGI predictions, DNA methylation data, cancer information, genes and mutations and correlation among DNA methylation, gene expression and cancer[13].

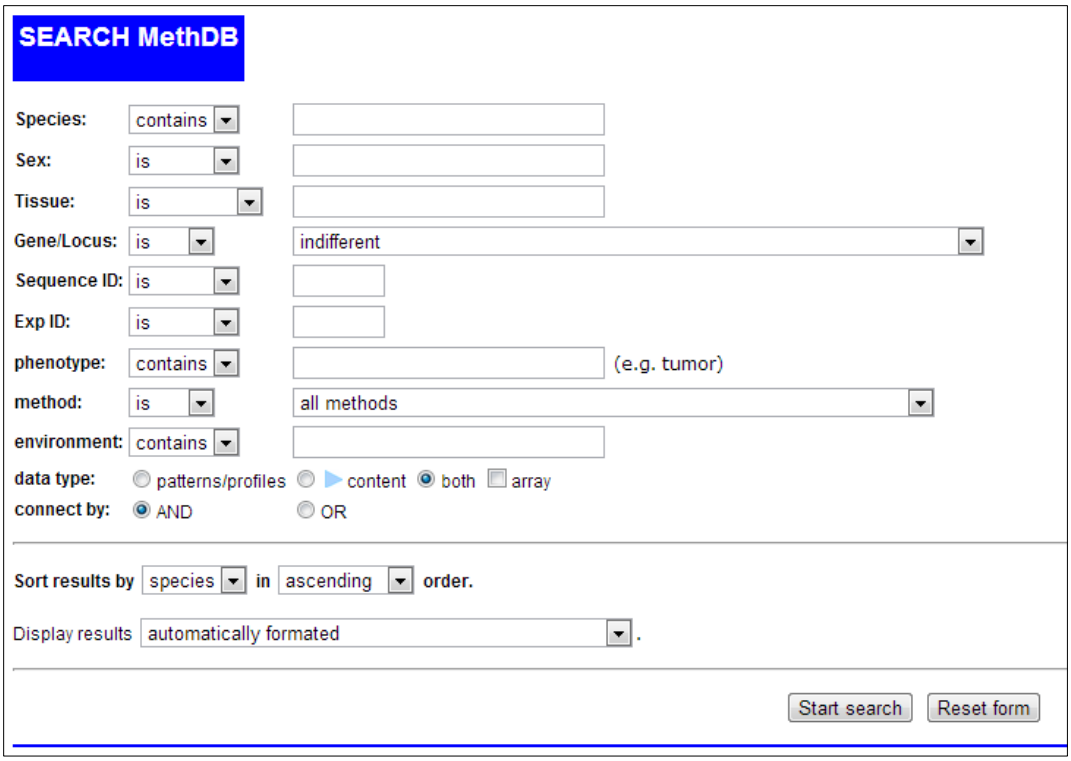


Figure 3 Search panel for MethDB database

DNA methylation data for the database is collected from MethDB, HEP and Methylation Landscape of Human Genome at Columbia University (Columbia)[14].

MethDB database stores information related to the degree of methylation in total DNA, DNA fragments and single nucleotide positions[15]. It also contains data on the nature and origin of the samples along with data obtained from the experiments. It provides a graphical and alphanumeric representation of methylation patterns and profiles helping in bringing the heterogeneous data on DNA methylation under one destination. The entries present in the database are cross-linked to other databases by including hyper-links to external data resources. Results received from database are distinguishable into 3 forms: methylation content, methylation profile and methylation pattern.

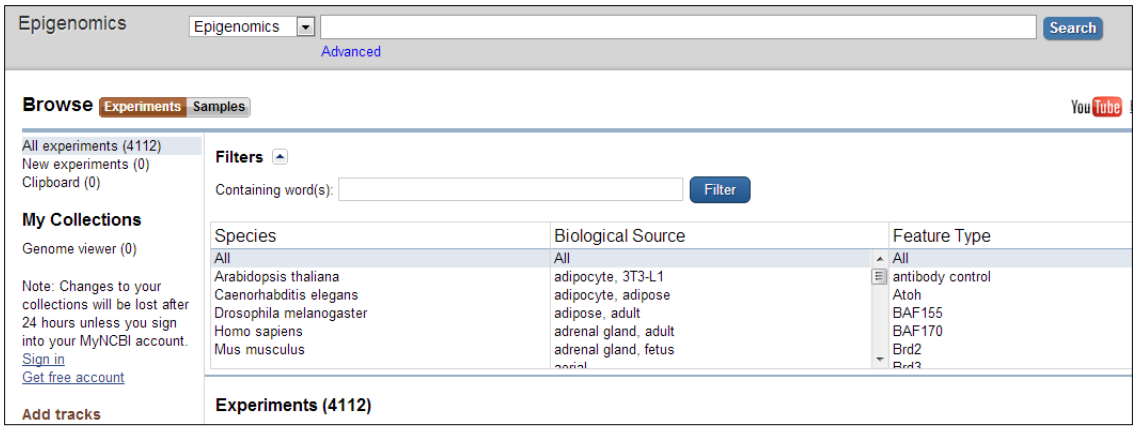


Figure 4 Search panel for Epigenomics database

NCBI's Epigenomics database is constructed by picking epigenetics oriented data from data archives like Gene Expression Omnibus and Sequence Read Archives[16]. Epigenetic data collected is then put to review, annotation and reformation. The raw data is processed by mapping it to generate genomic coordinates. These genomic coordinates are used to build genomic tracks which are used for visual representation of the data. The coverage of Epigenomics database currently covers data tracks for DNA methylation, histone modification, expression of small non-coding RNAs and chromatin accessibility. Transcription factors, components of the core machinery and histone modifying enzymes which are considered as chromatin associated factors are also made available through the database. 'Studies' and the 'samples' are two basic types of records in Epigenomics database.

Even though the above mentioned databases cover important epigenetic data, they fail to provide a universal view of epigenomics. PubMeth database does not cover all genes responsible for epigenetic modifications. Text-mining sources for the database are restricted to the abstracts from PubMed and Medline which does not include the complete set -+of literature that is present for epigenetics. Apart from text-mining related

statistics the database has no extra information related to epigenetics associated genes. It also fails to provide quantitative measures for DNA methylation levels. MethDB only focuses on DNA methylation whose information is general and sample oriented. In addition to this it is not optimized to cancer related queries either. MethyCancer also as an epigenetic database fails to answer every aspect of epigenomics modification. The data present in the database is not updated and the interface provides a complicated view which makes it difficult to query the database. NCBI's Epigenomics primary data sources are GEO and SRA which inhibits database to contain complete data related to epigenetics. Considering the shortcomings and pitfalls of current epigenetic data resources there is a need for new data reserve which will have a greater coverage of the data and has information for better understanding the field.

2.3 Thesis Statement

Currently all major epigenetic databases lack in terms of the coverage of data stored in them and further these databases fail to undertake a universal data processing methodology. Each database has its own data processing pipeline which mostly focuses on cell lines rather than cancer and their sub-types as whole. They lack information which would help to provide a full epigenetic view for given components of the machinery. Most of the databases also fail to provide information on environmental factors which play a pivotal role in defining the epigenomics landscape. There is also a poor mapping or association between epigenetically affected genes and their role in major cancer pathways. Considering these shortcomings led to the idea of building a standard pipeline for analyzing epigenetic data and building a robust but universal data model which will house the data processed from this pipeline.

This thesis mainly focuses on DNA methylation aspect of epigenetics as it has the major contribution in observed and cancerous epigenetic aberrations. We have built a

robust pipeline for processing the DNA methylation data obtained from 'The Cancer Genome Atlas (TCGA)' (http://cancergenome.nih.gov/). The pipeline processes the methylation data for different patients from TCGA using various statistical concepts and combines system biology[17] to make meaningful sense from the results obtained. Each cancer methylation data is processed stage-wise so the pipeline provides granularity in understanding the process as well helps to have a high level comparative view to understand the progression of methylation across different stages of cancer. Finally the results obtained from this pipeline are dumped into a database modeled to hold information for each cancer, stage-wise, and provide epigenetic attributes such as environmental factors, pathways, protein-protein interactions which help to understand the methylation state of a gene and its effect on other important cellular mechanisms.

The scope and importance of pipeline is proved by using 'Lung Adenocarcinoma (LUAD)' as a case study. Lung cancer is one of the most common cancers. In United States 226,160 new cases were expected to be diagnosed in 2012 (http://www.cancer.gov/cancertopics/types/lung). Lung cancer is morphologically divided into non-small cell (NSCLC) and small cell (SCLC)[18]. Lung adenocarcinoma (LUAD) is currently the most common of the lung cancers in both smokers and non-smokers. We have analyzed the DNA methylation data for different patients from TCGA repository and have provided a stage-wise result set. Using system's biology we have obtained methylation patterns for LUAD across its different stages. These results are stored along with the stage-wise results for other cancers in our epigenetic database. Our database helps researchers to make epigenetic sense from the data made available to them.

Chapter 3 Methods

3.1 Analysis Pipeline

In order to understand a methylation data for a particular cancer in stage specific manner, the analysis pipeline was divided into following sections:

Section1. Identification of significant genes from expression data

The level 3 data available from TCGA was segregated based on the stages provided in Metadata. If 65% of the data for a gene was $\log_2 \geq 1.4$ or ≤ -1.4 , then the gene was considered for further analysis as it obeyed the stringency with respect to fold change > 2.5 (a $\log_2$ ratio of 1 represents a 2-fold change) [19]. The average value for each gene was then computed and considered for the next level analysis. If a gene was represented by two or more probes, then the median of its expression value was used.

Section2. Identification of significant DNA Methylated genes from methylation data

The beta-values [20], for normal and disease samples were downloaded from the TCGA for Illumnia HumanMethylation27 and stratified by stage. The difference between the normal and the disease beta-values were then calculated. Genes with beta-values greater than 0.25 were considered hypermethylated and those with beta-values less than -0.25 were considered hypomethylated [20]. Using the Mann-Whitney U test [21], p-values were computed for each gene. This test was considered as it can handle variance for unequal sample sizes. For the study the analysis of q-value and 1% FDR gave threshold for the p-values for all the stages[22]. Our analysis identified p-value < 0.001, as the optimal threshold across all the stages. These genes were termed as “Significant DNA methylated genes”. Since the sample sizes were small, to get true

inferences resampling technique was performed. The samples were permuted large number of times (1000) and Mann Whitney Test was performed for each permuted samples and p-values were computed[23].

The DNA methylated genes were then annotated with pathway information using KEGG [24]. This information was used to understand the stage-wise profile of pathways consisting of DNA methylated genes. This analysis found common and unique pathways across stages. One of the limitations of the pathway analysis was that the analysis was limited to the mapping of DNA methylated genes with KEGG pathways.

Section3. Understanding the DNA methylated genes based on stage-specific networks

To understand the significance of the DNA methylated genes for a cancer, stage-specific networks were obtained using the following steps:

Identification of gene-gene interactions and DNA methylated-gene interactions from BioGRID

The gene-gene physical association among significant genes and DNA methylated genes was identified using BioGRID [25]. This integrated network was analyzed across different stages to capture the differences and commonality based on the following criteria: (i) two DNA methylated genes interacted; (ii) the DNA methylated gene has interaction with an expressed gene; or, (iii) the DNA methylated gene has interactions with a gene other than the significant genes in the given stage. This association was termed as “missing link” and the gene as “novel gene”, if the novel gene has an interaction with an expressed gene in the given stage or in other stages. Also the novel gene was expressed in previous or subsequent stages. This novel gene was then

evaluated using biomedical literature for its significance with the DNA methylated gene and the given cancer.

The nodes and the edges of each stage-specific network were annotated with their respective topological and biological features. The statistical computing tool R (www.r-project.org) was used to compute the topological features betweeness and clustering coefficient for each node (gene) in the network. The two biological features considered for analysis were: Pathway Significance Score and Gene Ontology Semantic Similarity. The Pathway Significance Score was based on the occurrence of the given gene in a pathway class i.e., the lung cancer pathways, other cancer pathways, or other pathways (includes metabolic) as given in KEGG [24]. These features were normalized individually and the average of these features defined the NodeStrength of a node (gene), given as:

$$NodeStrength_v = \frac{(Betweenness + Clustering\ coefficient + Pathway\ significance\ score)}{3} \quad \text{(i)}$$

Betweeness of a gene v was defined as the inverse of the ratio of the total number of shortest paths from gene s to node t given by σ_{st} to the number of total paths passing through node $v(\sigma_{st}(v))$ [26]. This was computed as:

$$Betweenness\ \left(B_{bet}\ (v)\right) = \sum_{s \neq v \neq t} \frac{\sigma_{st}\ (v)}{\sigma_{st}} \quad \text{(ii)}$$

Clustering Coefficient (C_v) was defined as a function based on the triplets of the genes in the network, where a triplet consisted of the three genes (nodes) connected by either

two open or three closed undirected ties [27]. The clustering coefficient for the genes in the undirected graph (stage- specific network) was computed as:

For a graph $G = (V, E)$ consisting of vertices V and a set of edges E, where $e_{i,j}$ connects vertex v_i vertex v_j and the neighborhood N_i for this vertex v_i was defined as:

$$N_i = \{v_j : e_{ij} \in E\}$$

(iii)

And where k_i represents the number of vertices in the neighborhood of N_i. The clustering coefficient for this local graph was then computed as:

$$Clustering\ coeficeint\ (C_v) = \frac{2\ |\{e_{jk}: v_j, v_k \in N_i, e_{jk} \in E\}|}{k_i\ (k_i - 1)}$$

(iv)

Pathway Significance Score

The pathways associated with each node's v (genes) were identified using KEGG, and Pathway Significance Score was computed as;

$$\text{Pathway Significance Score}_{\text{v}} = \left\{\log_{10}\left[\left(\frac{\text{frequency of term}}{\text{Total frequency}}\right) * 100 * (\text{Total frequency})^{\text{strength}}\right]\right\}^{\text{strength}}$$

(v)

Where, Pathway Significance Score for lung cancer case study determined the level of importance of a gene in the lung cancer pathways, non- lung cancer pathways (can be other cancer pathways) and other pathways (i.e. pathways that are not termed as lung

cancer pathways or non-lung cancer pathways); frequency of terms equaled the count of the gene in lung cancer pathways, non-lung cancer pathways and other pathways; Total frequency was equal to the count of the lung cancer pathways, non-lung cancer pathways and other pathways; Strength represented the importance of the pathway in the stage-wise network. For all the stage-specific network lung cancer pathway was given a prior score of 3, non-lung cancer pathways were given a prior score of 2 and other pathways were given a prior score of 1.

EdgeStrength

For any two interacting nodes (genes) in the network, EdgeStrength was computed based on their Gene Ontology Semantic Similarity. This was calculated using the GOSemSim package R [28].

All the genes and their edges in the stage-specific network were then annotated with their NodeStrength and EdgeStrength. The DNA methylated genes were ranked based on their NodeStrength. The highly ranked DNA methylated genes were used to identify subnetworks as described in the following section.

Section4. Identification and scoring of epigenetically relevant subnetworks across stages

To understand the functional significance of DNA methylated genes in the different stages of cancer, graph techniques were used to identify the subnetworks [29-31]. To compare and elucidate the interaction network of DNA methylated genes across stages is a hard problem. Therefore in this work, subnetworks of different sizes were identified and analyzed across the stages to understand the interaction profile of DNA methylated genes. As these were open subnetworks i.e. no size and shape limitation, therefore an NP-hard problem. To understand these subnetworks functionally, the genes

in these subnetworks, were analyzed based on KEGG pathways and these were understood in four categories: (i) genes identified in cancer pathways other than lung cancer pathways, (ii) genes identified in lung cancer pathways, (iii) genes identified in signaling pathways (not present in (i) and (ii), and (iv) genes in the metabolic pathways and other pathways. Starting with the DNA methylated gene as a seed, its interactions were identified, propagated, and analyzed based on the above four different categories. These subnetworks correlate to distinct functions that specify the distinct mechanism that can be compared across the stages. These subnetworks were further analyzed identified based on their NodeStrength and EdgeStrength.

The DNA methylated genes in each stage were ranked based on their beta-value. The DNA methylated gene with the highest beta-value was considered as a SEED. The SEED and expand algorithm was then used to identify the next connecting gene and edge based on the NodeStrength and EdgeStrength. Thus, subnetworks of different sizes were identified and connected in each of the stage-specific network and scored based on their SubnetworkStrength which was computed as;

$$SubnetworkStrength = \frac{\sum_{i=1}^{i=k}(NodeStrength) + \sum_{j=1}^{j=k-1} EdgeStrength}{Number\ of\ Genes} \quad \text{(vi)}$$

Where, i are nodes, j are edges, and k is number of nodes or edges.

As this is an open network and subsequently an NP-hard problem, as large number of subnetworks of different size are possible. Analyzing these subnetworks individually with respect to all its nodes (genes) is both time consuming and hard. Therefore, the subnetworks were further classified based on the nodes (genes) that were present in the different pathway class, namely, Lung Cancer pathways, Cancer

Pathways, Signaling Pathways and Metabolic Pathways. The subnetworks were compared for their commonality and uniqueness across stages to identify the possible DNA methylated genes that could be potential targets. The literature was then consulted to validate the significance of these DNA methylated genes.

3.2 Data model for Epigenetics Database

The data model for epigenetic database was designed based on the requirement to hold the results from DNA methylation data processing pipeline and the data obtained from different web resources which help to define the epigenetics characteristics of a significantly methylated gene. The data model shown in Figure 5 is designed using MYSQL workbench[32].

For analysis results the model holds information such as gene symbol, beta-value for showing the methylation level of gene in a particular cancer, p-value to determine the significance level of the beta-value and stage-wise information. The data model also supports information from other important databases such as KEGG database[24], some tables from UCSC Genome Browser[33], BioGRID database[25] and GO database. KEGG database provide information on the pathways and the disease annotation for these pathways along with the information on the environmental factors affecting the pathway[24]. This relation was of special importance to the model as it helped us to make a connection between the significantly methylated genes, pathways they are involved in and the possible environmental factors responsible for the current state of genes in the cancer condition. The association helps to distinguish our data resource from the present epigenetic data resources. UCSC Genome Browser tables help to provide annotations and gene symbol conversions between various data sources used in the model[33]. The ID conversion provided by the Genome Browser act

as a foreign key between multiple important tables in our database. Whereas BioGRID fill in the physical interaction information which helps us to understand the role of significantly methylated genes in larger cellular mechanisms and points us to the possible area of the cell that needs to be analyzed for better understanding cancer[25].

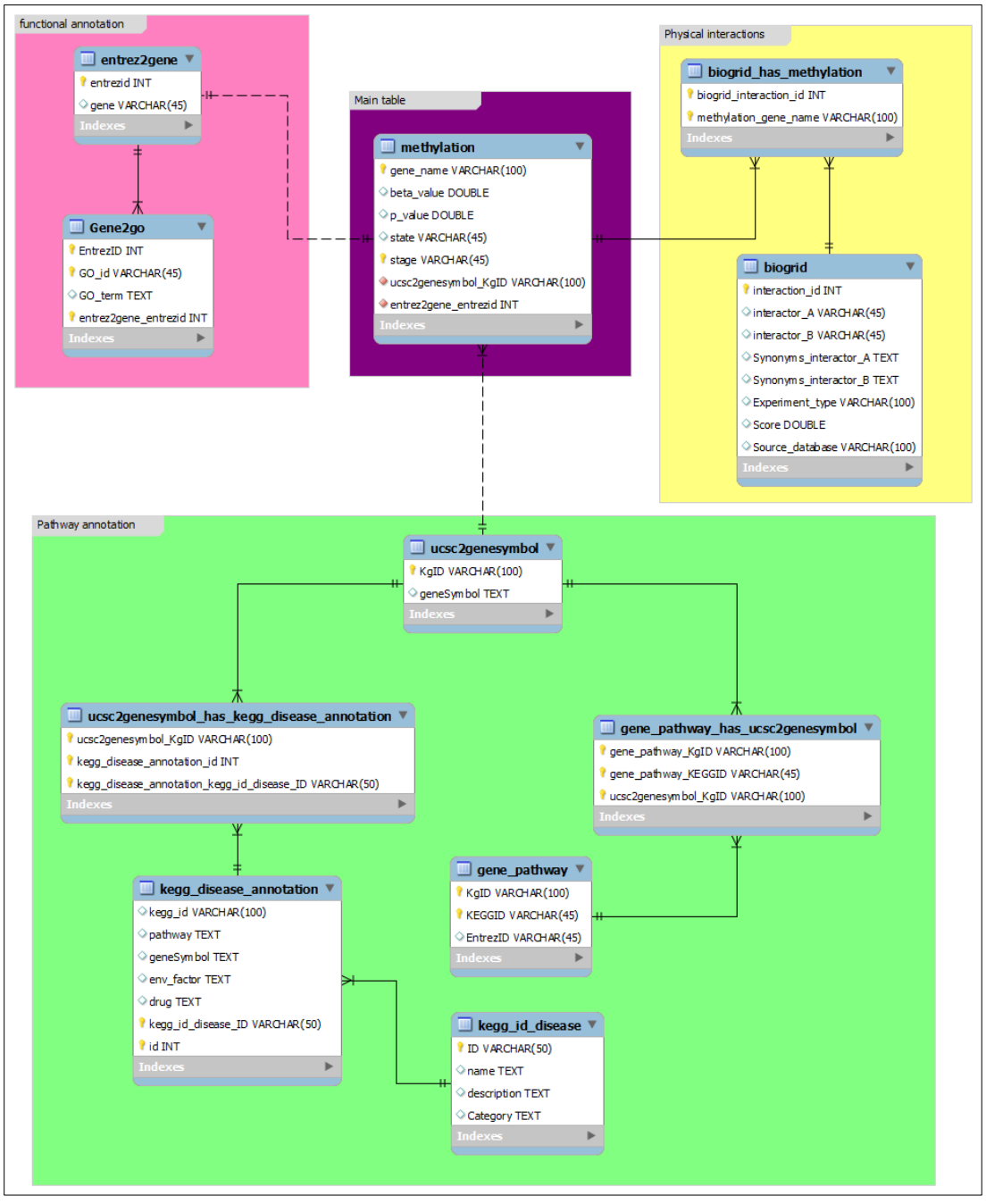


Figure 5 Data model for Epigenetics database

3.3 Implementation of Data model for Epigenetics Database

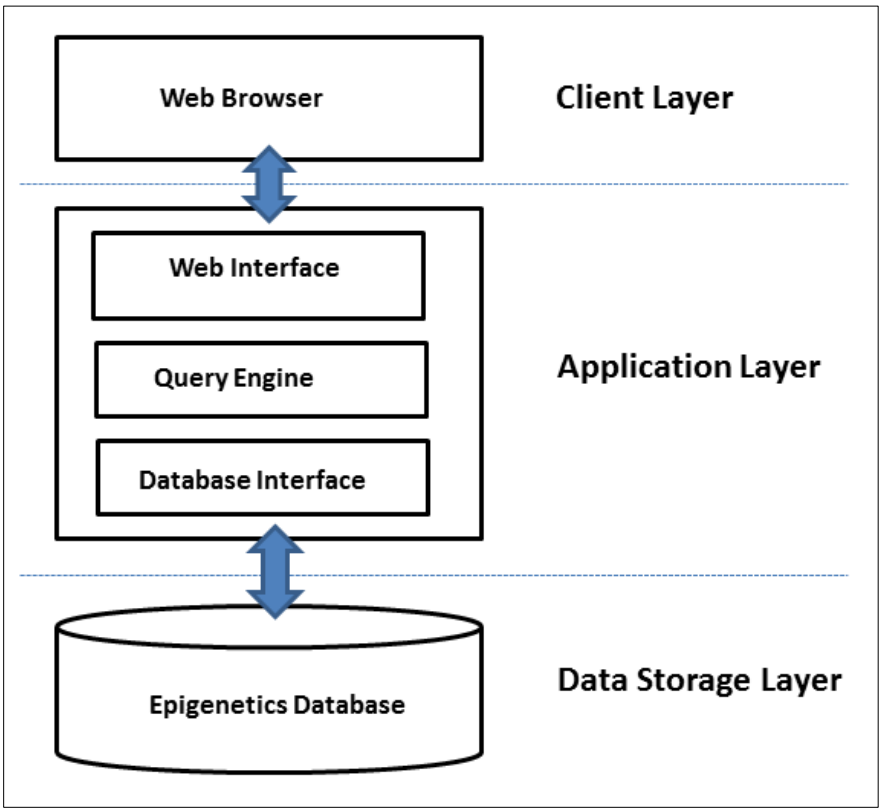


Figure 6 Implementation of Database

The data model for Epigenetics Database is implemented in 3 layers: Data storage layer, Application layer and Client Layer as shown in Figure8. To implement the architecture Linux Apache MySQL PHP (LAMP) stack was used.

Data Storage Layer:

Database was implemented using MySQL database management system. The database has two interfaces. At first interface the data is dumped into the database as shown in Figure6 while at the second interface the data is pulled out of database so that it can be displayed by the web application. The query for inserting data in database is:

INSERT INTO methylation (gene_name,beta_value,p_value,state,stage,cancer,platform)

VALUES (**from Python Parser**);

The data from other online data resources was directly imported from text files downloaded from these databases. The command used for importing the data from tab delimited text files:

LOAD DATA INFILE 'database.txt' INTO TABLE epigentics_database_table FIELDS TERMINATED BY '\t' LINES TERMINATED BY '\n';

Application Layer:

The application layer has 3 main components: Database Interface, Query Engine and Web Interface. Database Interface was implemented by forming a connection between PHP and database server. In order to make this connection “mysqli_connect ()” function from standard PHP library was used by providing proper credentials and the database name. Query Engine includes a collection of SQL queries that were used to extract relevant information from the database to fulfill various functionalities that are made available through interface. Web interface for the application was built using PHP and the site is hosted on Linux based server. Additional concepts of web programming were used for web interface implementation. Cascading Style Sheets (CSS) were used for designing various elements used on web page. Complex logic was implemented by making AJAX calls to various web pages. The whole of data storage and application layer is present on a server named “Regen” belonging to TiMAP group at IUPUI.

Client layer:

Client layer involves usage of browsers such as Mozilla Firefox, Google Chrome, and Internet Explorer by users to access the web interface from application layer. The link to access our web interface:

http://regen.informatics.iupui.edu/cgi-bin/epigenetics/DEVenv/index.php

Through this layer users can select different options that are made available to them through web interface. These options then get converted into queries and fetch corresponding data from database. These actions are processed by initiating a HTTP requests to the server and completed by receiving an appropriate response from web server. Apache server application was used to handle HTTP requests from client browser, process them appropriately and to respond with a proper HTTP response.

3.4 Python Parser

In order to implement the above mentioned DNA methylation data analysis pipeline and integrate the results from pipeline to the data model of database, a python parser was built to carry out the necessary functionalities.

The parser has 3 main roles to perform: collect raw data for respective cancer type from TCGA data repository, process the collected data through the designed pipeline and finally dump the results in the database in the format set by data model. A user provides a link for raw data to the parser. After the link is provided the parser downloads datasets for the concerned cancer. Along with the link to the raw datasets a metadata file containing the stage-wise information for the patients whose data is downloaded needs to be provided. Following the downloading step, the parser parses the metadata file and segregates the methylation data into different folders according to

the stages defined in the file. It generates a tree structure for the files to be stored. The topmost node of the folder structure is named with the cancer name. This node is then sub-branched into the folders with stage number on them. The automation of this step saves the user from the laborious task of manually segregating the dataset files. For segregation the parser uses dictionary data structure in which the stage number is stored as a key and patient data filenames as values. When it scans the super set of datasets it uses this data structure to correctly divide the files in respective stage folders. Once the data is segregated correctly, the next step is of data cleaning and merging the datasets. At the data cleaning step duplicates are handled in the file by calculating median between multiple beta-values for same gene. Once the file is remedied from duplicated beta values the data for individual patients is merged and again the folder tree structure is further sub-branched into normal and cancer folders. This sub-branching happens for each stage present in the cancer.

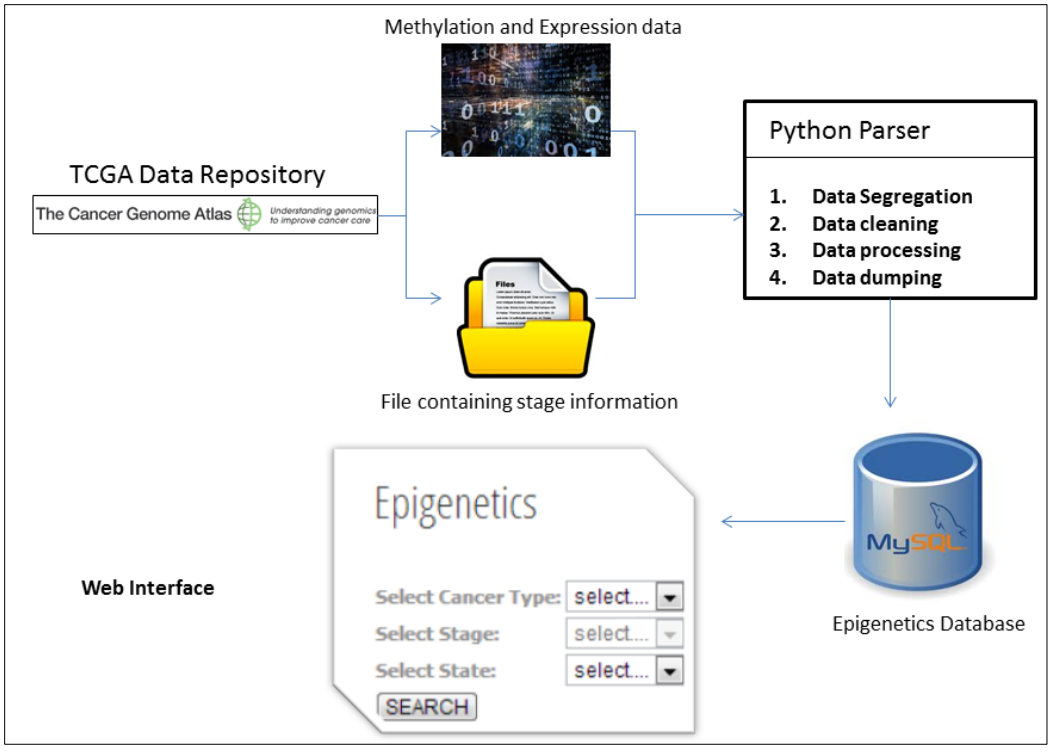


Figure 7 Python Parser

The data is now ready for identifying significant methylated epigenetics genes for the cancer. The statistical processing on the dataset was performed by using the appropriate statistical packages from R. In order to integrate python parser with R environment, system calls are made to the R interpreter from Python code. R provides a wide range of flexibility in setting different parameters for the statistical functions present in it. Once the analytical steps are performed on the data, the data is molded in a template that is consistent with the data model used to store data in epigenetics database. As soon as the data is loaded in database and a positive token is received from database server, the lifecycle for Python parser is finished.

3.5 Case Study: LUAD

The overall methodology for the stage-wise identification of LUAD process is shown in Figure 8 that includes four unique steps (A-D).

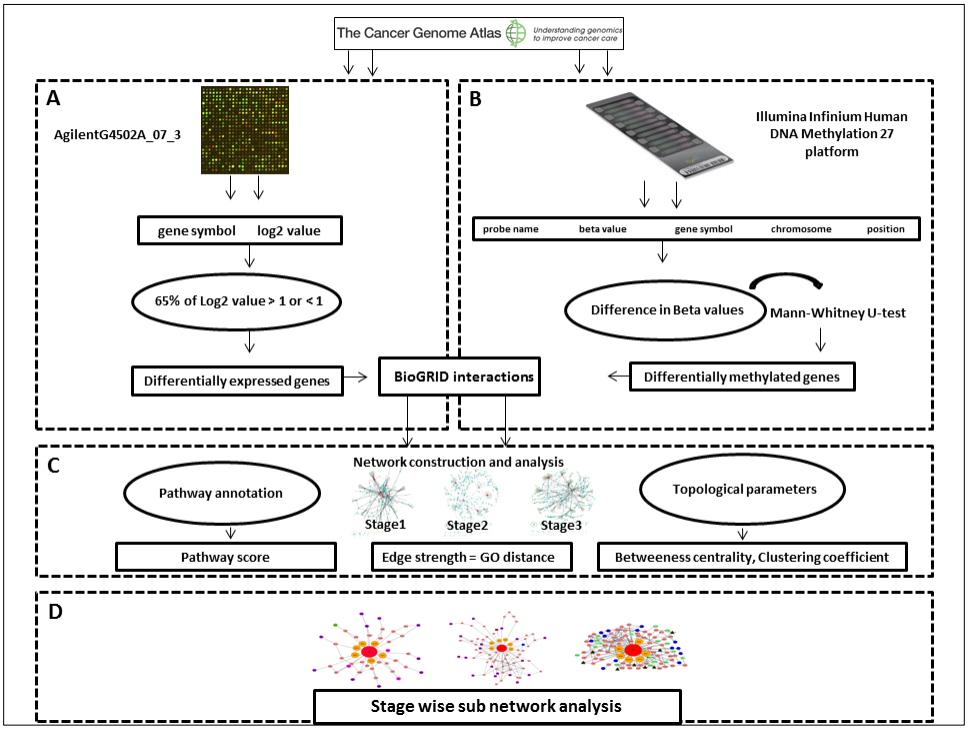


Figure 8 Methodology to obtain patterns across stages of LUAD

Step A: the gene expression data from UNC AgilentG4502A_07_3 was analyzed based on the log2 values to obtain the differentially expressed genes.

Step B: the methylation data from Illumina HumanMethylation27 for each stage was analyzed based on the beta value to obtain the differentially expressed methylated genes.

Step C: the data obtained from step A and B was integrated to obtain a stage-specific network of LUAD. This network was annotated with the topological and biological features and analyzed for methylated pattern identification.

Step D: the stage-specific subnetworks were obtained for LUAD. Details of each of these steps are now described in the following sections.

Details of each of these steps are provided in the analysis pipeline section.

Data

The gene expression and DNA methylation data for LUAD were downloaded from TCGA. The gene expression data were generated by UNC AgilentG4502A_07_3 and the methylation profiles were generated by Illumina HumanMethylation27 DNA Analysis which contains 27,578 CpG dinucleotides in 14,495 genes[34]. These dataset were downloaded on 10-12-2012 and arranged with respect to the four stages of LUAD. Across all these stage-wise data, the patient's age ranged from 58-75 with few outliers.

Chapter 4 Results

The objective of this thesis was to design DNA methylation data analysis pipeline and prove the robustness of database as a universal epigenetic analysis result storage unit. The database currently holds epigenetically relevant information for: Lung Adenocarcinoma (LUAD), Colon Adenocarcinoma (COAD) and Breast invasive carcinoma (BRCA). Each cancer has the significant methylated gene information segregated in stage-specific manner. The stage specific methylated gene is further annotated with data integrated from various other databases to make epigenetic sense out of data. Table 12 in Appendix section shows the fields present in databases selected for annotations. The table also states reason for selecting particular database for annotation of methylated gene. LUAD was considered as a case study to prove the effectiveness of pipeline in analyzing DNA methylation data sets to obtain meaningful analysis. The TCGA data associated with LUAD was classified based on four stages. Across all these stage-wise data, the patient's age ranged from 58-75 with few outliers.

4.1 Effectiveness of analysis pipeline

Identification of highly scored significant DNA methylated genes

The Significant expressed genes and Significant DNA methylated genes were identified based on the p-values and beta-values for each stage as described in the methodology. Resampling technique performed the correction and provided the set of p-values. Using the technique used in paper [22], p-value of 0.0012 was obtained from q-values. Using this cutoff the Significant DNA methylated genes were re-evaluated and overlap between the previous and resampled results was calculated. A substantial amount of overlap between old and new set of Significant DNA methylated genes was observed. Figure 16 in Appendix shows the p-value correction for original and corrected

Stage I data after resampling. The DNA methylated genes were then further classified as hypermethylated and hypomethylated (methodology section). Table 8 lists the statistics for each stage.

The significant DNA methylated genes were analyzed and ranked based on their beta-values. Table 2 lists the top 10 hyper/hypomethylated genes across stages in descending order of their beta-values.

As shown in Table 2, 10 of the highly methylated genes in Stage I were common across the three Stages (Table 1). Of these 10, 7: AJAP1, ATP8A2, HOXA9, PTGDR, SIX6, TLX3, TMEM130 were hypermethylated and the three: KRTAP8-1, MMP26 and REG3A were hypomethylated. Three of the seven (Stage I) genes: AJAP1, TLX3, PTGDR were also identified in Stage III. Interestingly the three top scored hypomethylated genes in Stage I were identified as top scored hypermethylated in Stage II. In addition, some of the top scored DNA methylated genes were common across two stages only (Table 1): LY96 was the top scored hypomethylated gene and top scored hypermethylated in Stage I and II respectively. While HOXA4, HOXD10, KRTAP15-1, LEP and NKX6-2 were identified as common across Stage II or III (Table 1). Table 2 also identified unique top scored DNA methylated genes i.e. 3 for Stage I, 8 for Stage II and 13 for Stage III. Comparison of highly DNA methylated genes across stages (Table 2) with common methylated genes across stages (Table 1) also depicts that a large number of the highly methylated genes are common across the stages and these can play an important role for understanding the disease.

Table 1 Common DNA methylated genes across stages

Stages	DNA methylated genes	
	Number	List of genes
Common DNA methylated genes across the three stages	34	AJAP1, ATP8A2, CCDC140, CNTP2, CYYR1, EVX1, FERD3L, FOXG1, GRIK3, GRM6, HAND2, HOXA9, HOXB4, HOXD4, HOXD9, HOXD12, INPP5B, OTX2, KRTAP8-1, MMP26, PHOX2A, PLEKHA6, POU4F2, PRAC, PRKCB, PTGDR, REG3A, SIX6, SLC6A2, SPAG6, TBX20, TLX3, TMEM130, ZNF560
Common DNA methylated genes across Stage I & II	12	ADCY4, BHMT, C12orf34, CDO1, LVRN, LY96, MSC, PCDHGA12, POU3F3, ZNF154, ZNF577, IHH
Common DNA methylated genes across Stage II & III	30	BARHL2, C10orf81, CCDC140, DEFB119, DIO3, FAM135B, FAM83A, GRIK2, HOXA4, HOXD10, HS3ST2, KCNS2, KRTAP15-1, LEP, LHX1, LYPD5, MAGEB6, NEUROG1, NKX6-2, OR5I1, SERPINB5, SPHKAP, TAL1, TBX4, TBX5, TCN1, TMEM132D, VSX1, ZNF454, IGKV7-3
Common hypermethylated genes across Stage I and Stage III	42	AJAP1, ATP8A2, CCDC140, CNTP2, CYYR1, EVX1, FERD3L, FOXG1, FOXI2, GALR1, GAS7, GRIK2, GRM6, HAND2, HLA-G, HOXA7, HOXA9, HOXB4, HOXD4, HOXD8, HOXD9, HOXD12, INPP5B, NID2, NPY, OTX2, PAX7, PHOX2A, PLEKHA6, POU4F2, PRAC, PRKCB, PTGDR, SIX6, SLC6A2, SOX17, SPAG6, TBX20, TLX3, TMEM130, VIPR2, ZNF560
Common hypomethylated genes across Stage I and Stage III	3	CORO6, MMP26, REG3A

Table 2 Identification of top beta-value scored DNA methylated genes across stages

Stage	Hyper/Hypo	Genes in descending order of beta-values (p<0.001)
I	Hyper	TLX3 > NEFM > PTGDR > AJAP1 > SIX6 > HOXA9 > TMEM130 > HISTIH3G > ATP8A2 > NID2
	Hypo	MMP26 > KRTAP 8-1 > REG3A > CORO6 > LY96
II	Hyper	LY96 > C10orf81> KRTAP8-1 > MMP26 > REG3A > DEFB 119 > NMUR2 > MAGEB6 > IGKV7-3 < KRTAP15-1
	Hypo	HTR2C > GRIK3 > CNTP2 > SPHKAP > TMEM132D > NEFH > LEP > ZNF177 > HOXD10 > NKX 6-2
III	Hyper	TLX3 > HOXB4 > AJAP1 > HOXA4 > HOXD9 > PTGDR > LYPD5 > FZD10 > HOXD12 > NECAB2
	Hypo	CHR6 > C13orf28 > TMEM156 > XDH > FGF6 > IVL > G6PC > KRTAP13-2 > C10orf39 > FCRL3

Network construction and analysis

A systems biology approach was used to determine the significance of the DNA methylated genes in terms of their associated expressed genes in each stage. The interactions between the significant genes (both methylated and non-methylated as listed in Table 3), were identified using BioGRID[25] and stage-specific networks were constructed. These networks were then annotated with respect to their DNA methylated genes.

Table 3 DNA methylated gene interactions across stages

Stage	Number of DNA methylated genes	Number of interactions
I	72	228
II	93	273
III	170	660

The gene interaction analysis showed that each DNA methylated gene also interacted with additional genes in BioGRID [25]. These additional genes were analyzed for their expression in all the stages to determine if DNA methylation affected their expression. These interactions were termed as “missing links”, and the additional genes as “novel genes”. Table 4 gives the profile of the missing links.

Table 4 Novel genes (Missing Link-methodology) discovered using BioGRID

Stage	Missing Links	Novel genes	Number of DNA methylated genes	Number of novel genes and the stages where these are identified
I	27	27	16	6(Stage II, III),6 (Stage II),3 (Stage III)
II	43	33	25	10 (Stage I, III), 2 (Stage III)
III	132	83	32	34 (Stage I, II), 3 (Stage I), 7 (Stage II)

Further analysis of the 27 novel Stage I genes for their significance in other stages indicated 6 of them in Stage II: ANXA7, APBB1IP, MDK, PFDN1, TINF2, TLE2; 3 in Stage III: CUL5, CTNNB1 and SQSTM1 and 6 in Stages II and III: CALM1, CTNNB1,c-

JUN, SMAD1, TINF2. Of the 33 novel Stage II genes 2 were associated in Stage III: A2M and CTNNB1I; and 10 genes in Stages I and III: FOXA2, HK3, NCF1, NRIP1, PDLIM1, SP1, SUMO1, TCF4, TLR4, and TNN. Analyses of the 83 novel Stage III genes found three in Stage I: ELN, FAS and TEX11; 7 in Stage II: ANXA7, APBB1IP, MDK, PFDN1, STAT3, TLE2, UBE2B and 34 in Stages I and II:BCR, DLG3m, DLG4, EGFR, DSP, MAFF, PICK1 etc.

Figure 9 shows the stage-specific networks of DNA methylated genes. From this figure it can be seen that Stage III networks were more connected and dense as compared to other two stage networks, suggesting the greater number of epigenetically modified genes in later stages can influence the LUAD network.

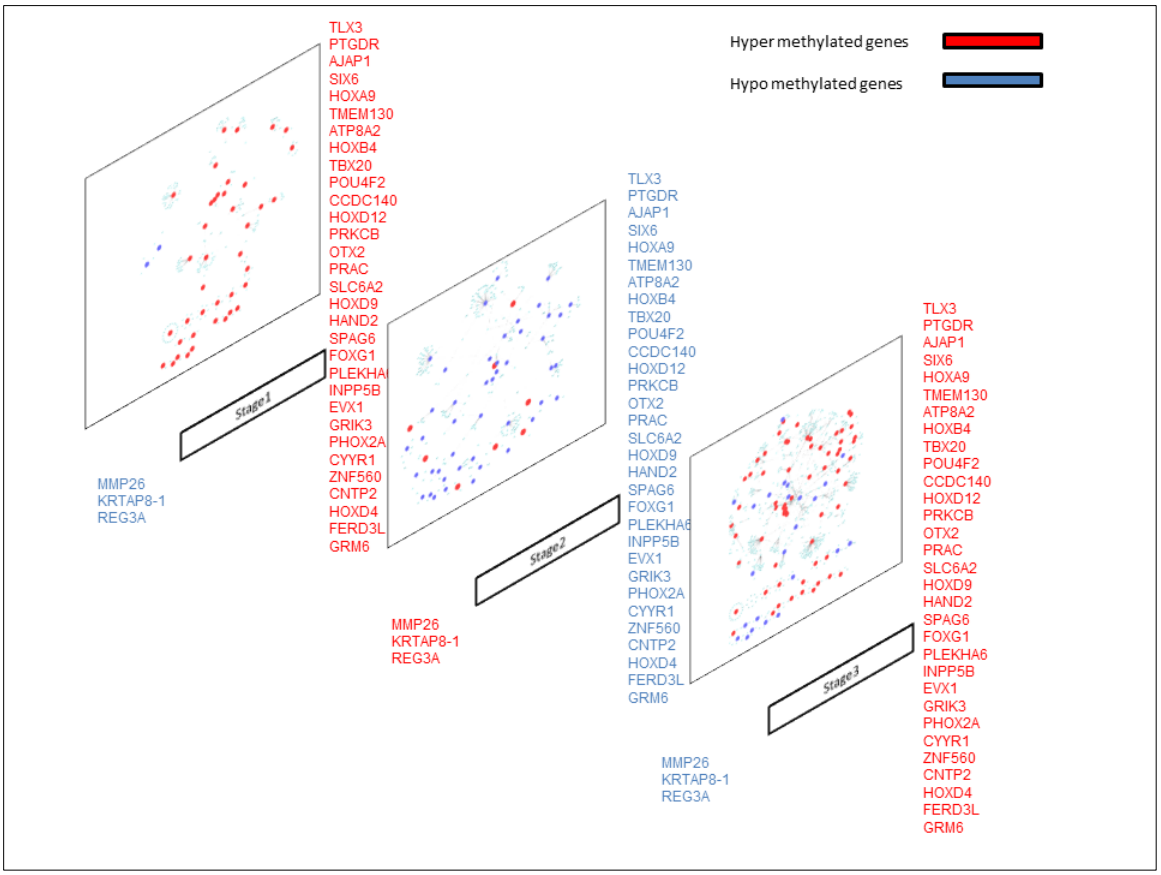


Figure 9 Stage-wise network patterns for methylated genes in LUAD

To compare stage-wise networks, DNA methylated gene subnetworks were identified and analyzed. Figure 15 in Appendix section shows the distribution of important classes of pathway across different stages of LUAD. SEED and expand algorithm (described in methods) was used to identify the subnetworks. Table 9 in Appendix list the number of subnetworks. These subnetworks were overlapping as the genes in them belonged to different pathway class. Table 9 in Appendix shows that the number of subnetworks drastically increases between sizes four and five in most of the stages, making it an NP-hard problem. This sharp increase in the number of sub-network suggests that though the DNA methylated gene is not directly connected to a hub node, its interaction path has a hub node. This further indicates that a DNA methylated gene can influence the whole network of a given stage. Table 5 lists the subnetworks with greater number of connections identified in all three stages.

Table 5 Analysis of hub genes in the DNA methylated subnetworks of size 4

Stage	Sub-network	Connectivity profile of the hub node across pathways			
		Cancer	**Lung cancer**	**Signaling**	**Meta-bolic+ others**
I	(i) PHOX2A*:HAND2*:PPP2R5D:UBC (ii) HLA-G*:COPB1:TRIM37:UBC (iii) LY96*:TLR4:SIGIRR:UBC	23	13	71	288
	(iv)HLA-G*:COPB1:UBC:CUL1 (v)FOXG1*:FOXH1:SMAD3:CUL1	21	14	35	114
	(vi)HLA-G*:COPB1:UBC:SKP2	19	10	17	38
II	(i)PHOX2A*:c-JUN:SUMO3:UBC (ii)TAL1*:HDAC1:IRF5:UBC	23	14	74	254
	(iii)PRKCB*:HIST1H3I:CUL4A:CUL1	20	14	74	253
III	(i)PHOX2A*:c-JUN:SUMO3:UBC (ii)PHOX2A*:HAND2*:PPP2R5D:UBC	22	10	64	235
	(iii)PRKCB:HIST1H3I:CUL4A*:CUL1	20	13	35	102

*: commonality with Table 1

As shown in Table 5, UBC and CUL1 were identified as hub gene across the three stages and their connectivity profile changes with pathway class. The other hub genes (number of connections) identified in Stages I and III were: SIRT7 (6), CDK2 (5),

PMS2 (4 connections), SUMO2 (3), SMAD3 (7), SMAD4 (5), and SMAD2 (4). The analysis also identified LY96 sub-network in Stage I consisting of the hub gene TLR4 interacting with seven other genes. Though LY96 was also identified in Stage II, the comparative sub-network was smaller and this gene was not identified at all in Stage III. HLA-G was present in Stage I but not in Stage II; therefore all its sub-networks were missing. In Stage II and III, c-Jun a TF was identified as a hub gene. PHOX2A was the DNA methylated gene associated with c-Jun in both stages. There was similarity across common gene (see Table 1) with Table 6, depicting that subnetworks constructed out of common genes across or between two stages can be of significance for LUAD. The size four subnetworks were further compared across the stages to understand their commonality and uniqueness (Table 10 in Appendix). These size four subnetworks were analyzed for their common DNA methylated genes. The common DNA methylated genes in these size four subnetworks present across cancer, lung cancer and signaling pathway classes were FOXG1 and PHOX2A (see in Table 1 also). The other common significant but not methylated genes present in these subnetworks were FOXH1, FOXO3, HAND2, MYC, RB1, SMAD2, SMAD3, SMAD4, and TP53. The unique genes in these subnetworks that were associated with other cancer and lung-cancer pathways were LEF1, AR, GATA4, and SKP2. Similarly, in the metabolic pathways class, the highly conserved common sub-networks of GRIK2, GRIK3, GRIK5, and GRID2 were identified across all stages. Of these GRIK3 was methylated in all the three stages (Table 1) and GRIK2 in Stages I and III (Table 1).

Analysis of these subnetworks is an NP-hard problem because these are large open subnetworks. To reduce the complexity, the nodes and edges were scored based on the NodeStrength and EdgeStrength as given in methods section. The top scored size four subnetworks of each stage (Table 5 and Table 9 in Appendix) were propagated

and compared to identify the largest conserved subnetworks across the stages. This analysis identified a sub-network of size 11 with seven conserved genes: UBC, KRAS, PIK3CA, PIK3R3, RAF1, BRAF, and RAP1A. The g:Profiler tool was used for the enrichment analysis on the top scored sub-network given in Table 6. This analysis showed that these subnetworks to be enriched with common genes across stages (shown in Table 1), indicating that commonality across stages of LUAD can be critical in identifying the target genes.

Table 6 Enrichment analysis of the top scored subnetworks

Stage	Biological Process	p-value	Genes in subnetworks
Common across all Stages	Activation of MAPKK activity	1.12E-03	PHOX2A*, HAND2*, PPP2R5D, UBC, KRAS, PIK3CA, PIK3R3, RAF1, BRAF, RAP1A
I & II	Co-SMAD binding	1.2E-05	FOXG1*, FOXH1, SMAD2, SMAD1, MED15, UBC, KRAS, PIK3CA, PIK3R3, RAF1, BRAF, RAP1A
II & III	Nerve growth factor receptor signaling pathways	2.21E-04	HOXD4*, INPP5B, SLC6A2, STX1A, VAMP1, UBC, KRAS, PIK3CA, PIK3R3, RAF1, BRAF, RAPIA
I & III	Positive regulation of peptidyl-serine phosphorylation	1.42E-03	NPY*, NPY1R, LSM7, NR1H2, RMI1, UBC, KRAS, PIK3CA, PIK3R3, RAF1, BRAF
I	Transmembrane receptor protein tyrosine kinase signaling pathway	1.87E-03	HLA-G*, COPB1, UBC, KRAS, PIK3CA, PIK3R3, RAF1, BRAF, RAP1A
II	DNA helicase complex	6.8E-05	SERPINB5*, UCHL5, ACTR8, ACTR5, UBC, KRAS, PIK3CA, PIK3R3, RAF1, BRAF
III	Nerve growth factor receptor signaling pathway	5.5E-04	HOXB4*, CREBBP, KLF13, UBC, KRAS, PIK3CA, PIK3R3, RAF1, BRAF, RAP1A

*: commonality with Table 1

4.2 Database statistics

In order to avoid the loss of methylation data due to statistical stringency, the data entering the database skips the q-value filter for the p-values. This allows a researcher to select a data filtration of his/her choice in downstream processing based on the p-values provided through the database interface. The statistics shown in table 7 varies from the statistics in table 8 for LUAD because of the q-value filtration.

Table 7 Distribution of methylated genes for cancer data in database

	LUAD		COAD`		BRCA	
	Hyper	Hypo	Hyper	Hypo	Hyper	Hypo
Stage1	134	21	507	441	0	0
Stage2	108	39	685	241	391	137
Stage3	231	119	616	340	208	124
Stage4	0	0	0	0	-	-

The methylated stage specific genes are annotated with following data:

1. Gene name
2. Beta value
3. P value
4. State
5. Stage
6. Pathway
7. Go Terms
8. Interaction
9. Disease

The disease column is further connected to “Environmental Factors” responsible for inception of the disease. This is an important link as the knowledge of “Environmental

Factors" is a key point in understanding the variation of methylation levels based on the external factors.

Table 8 Distribution of significant genes and methylated genes across the four stages of LUAD

Stage	Number of Normal Samples	Number of Disease Samples	Significant Genes	DNA Methylated Genes	
				Hyper	Hypo
I	9	35	15994	67	5
II	7	14	16275	20	73
III	5	11	14688	110	60
IV	2	6	14814	0	0

4.3 Database Interface

Epigenetics database can be queried using a combination of 3 options that are presented at the homepage of the interface for the database. Figure 10 shows the home page of the website for the database. The left panel on the page has external links to additional electronic resources of epigenetic data. Central panel holds the searching capability. First the user can select from the list of cancer types that are currently present in the database. Based on the cancer type associated stages get populated into the "Select Stage" dropdown. From the options displayed, one can select information for individual stages or can call for a combined result set of all stages present in the selected cancer. The third option provides the usability of filtering the result set on the basis of state of the gene whether it is hyper or hypo methylated. User can also choose to obtain a combined result set of hyper/hypomethylated genes for a cancer type and its stage.

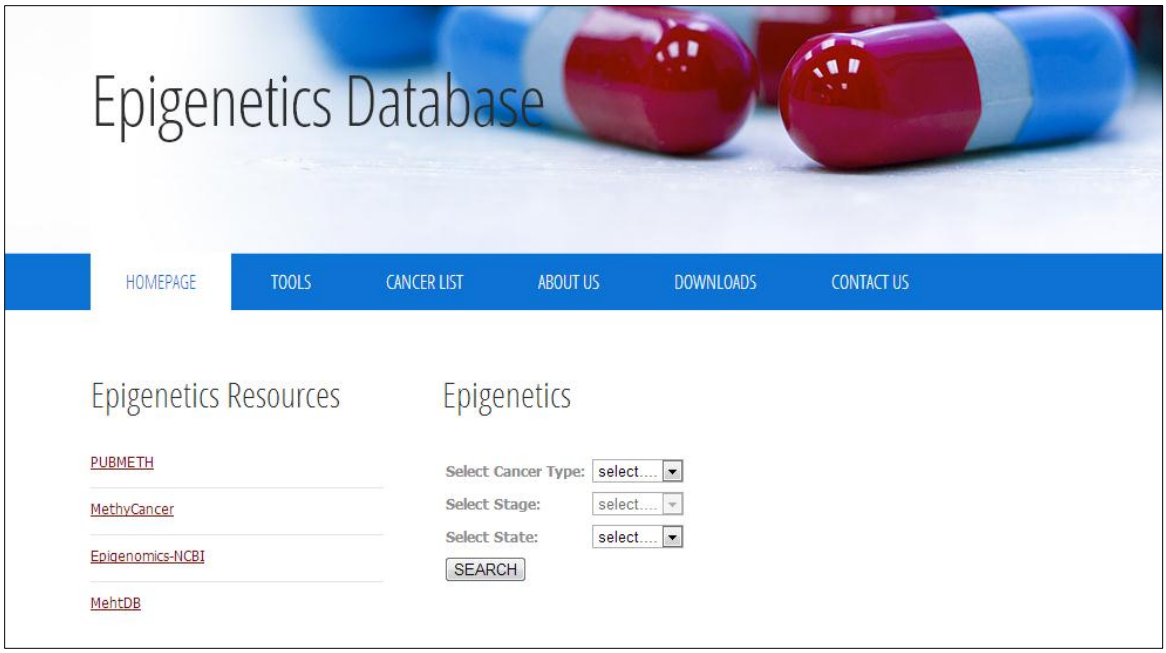
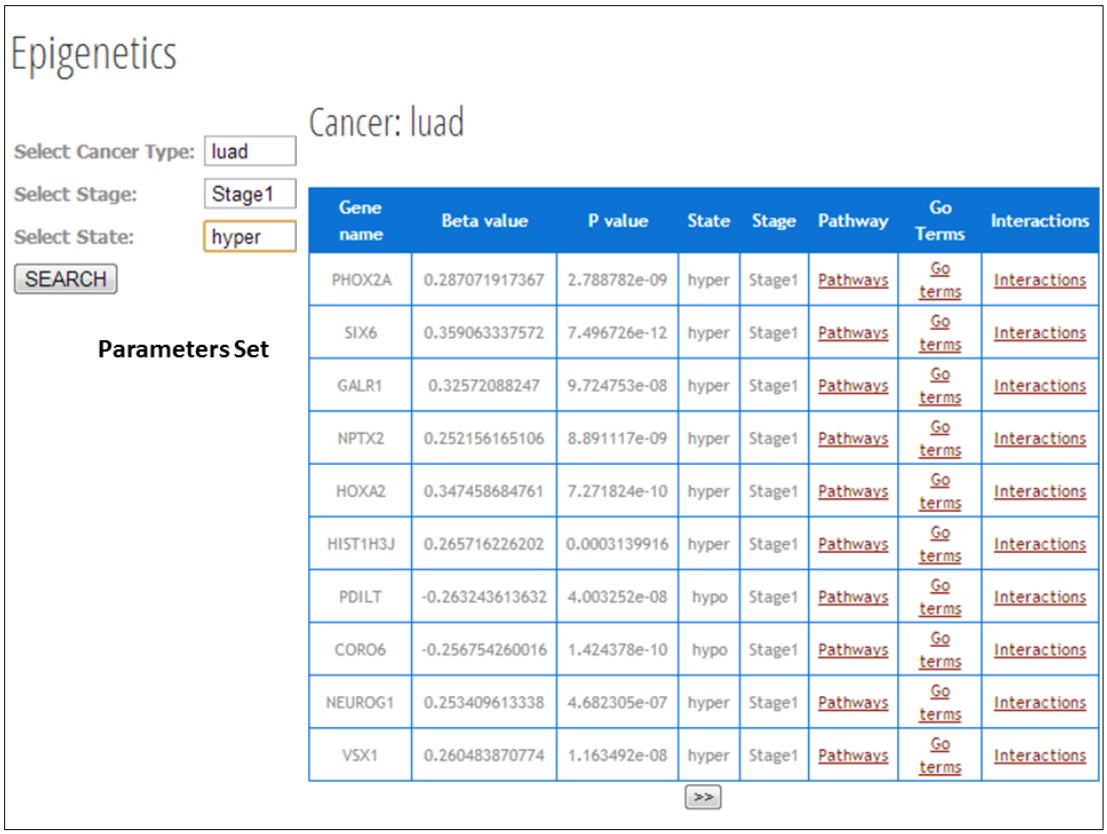


Figure 10 Home page for Epigenetics Database

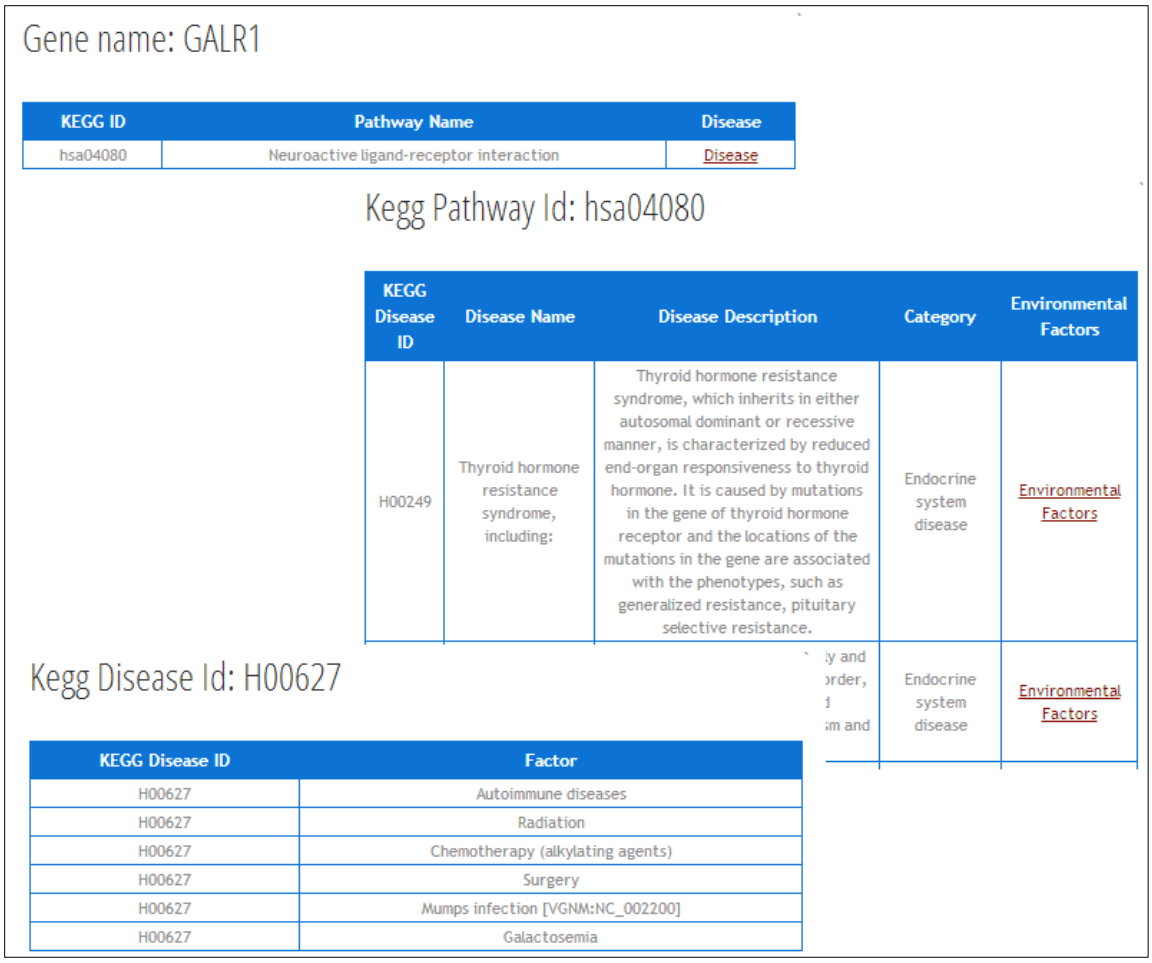
Figure 11 shows the LUAD result table filtered for hypermethylated genes in Stage 1 of cancer. Each gene in the table is characterized both statistically and with epigenetic relevant information. The statistics includes beta and p values for a methylated gene. These measures can be used by researchers to set up their own downstream processing experiments and derive results or networks for their respective researches. Beta values also state the methylation level of the genes in that specific cancer whereas the p-values state the statistical significance of the methylation level which is based on Mann Whitney test. Other epigenetically associated parameters such as pathways, GO terms, interactions and disease information help in understanding the role and nature of gene which can be used to make inferences for treating them.



Gene name	Beta value	P value	State	Stage	Pathway	Go Terms	Interactions
PHOX2A	0.287071917367	2.788782e-09	hyper	Stage1	Pathways	Go terms	Interactions
SIX6	0.359063337572	7.496726e-12	hyper	Stage1	Pathways	Go terms	Interactions
GALR1	0.32572088247	9.724753e-08	hyper	Stage1	Pathways	Go terms	Interactions
NPTX2	0.252156165106	8.891117e-09	hyper	Stage1	Pathways	Go terms	Interactions
HOXA2	0.347458684761	7.271824e-10	hyper	Stage1	Pathways	Go terms	Interactions
HIST1H3J	0.265716226202	0.0003139916	hyper	Stage1	Pathways	Go terms	Interactions
PDILT	-0.263243613632	4.003252e-08	hypo	Stage1	Pathways	Go terms	Interactions
CORO6	-0.256754260016	1.424378e-10	hypo	Stage1	Pathways	Go terms	Interactions
NEUROG1	0.253409613338	4.682305e-07	hyper	Stage1	Pathways	Go terms	Interactions
VSX1	0.260483870774	1.163492e-08	hyper	Stage1	Pathways	Go terms	Interactions

Figure 11 Result Table for options selected for query engine

The most important association of the database is shown in Figure 12. It shows the relationship between a gene, pathways it is present in and the disease caused by its malfunctioning. Disease field is then related to the environmental factors that may be responsible for the aberration of a gene's normal function. This loop completes the basic requirement of an epigenetic study which is the role of environmental factors in controlling the activity of a gene. Finally Figure 13 and 14 show the tables with GO terms and physical interactions for a gene. These tables help to elucidate the functions and interacting genes that may be affected by the methylation of a gene and can provide a local target for understanding the molecular working of a cancer.



Gene name: GALR1

KEGG ID	Pathway Name	Disease
hsa04080	Neuroactive ligand-receptor interaction	Disease

Kegg Pathway Id: hsa04080

KEGG Disease ID	Disease Name	Disease Description	Category	Environmental Factors
H00249	Thyroid hormone resistance syndrome, including:	Thyroid hormone resistance syndrome, which inherits in either autosomal dominant or recessive manner, is characterized by reduced end-organ responsiveness to thyroid hormone. It is caused by mutations in the gene of thyroid hormone receptor and the locations of the mutations in the gene are associated with the phenotypes, such as generalized resistance, pituitary selective resistance.	Endocrine system disease	Environmental Factors
		ly and order, l m and	Endocrine system disease	Environmental Factors

Kegg Disease Id: H00627

KEGG Disease ID	Factor
H00627	Autoimmune diseases
H00627	Radiation
H00627	Chemotherapy (alkylating agents)
H00627	Surgery
H00627	Mumps infection [VGNM:NC_002200]
H00627	Galactosemia

Figure 12 Relationship between gene, pathway, disease and factors causing the disease

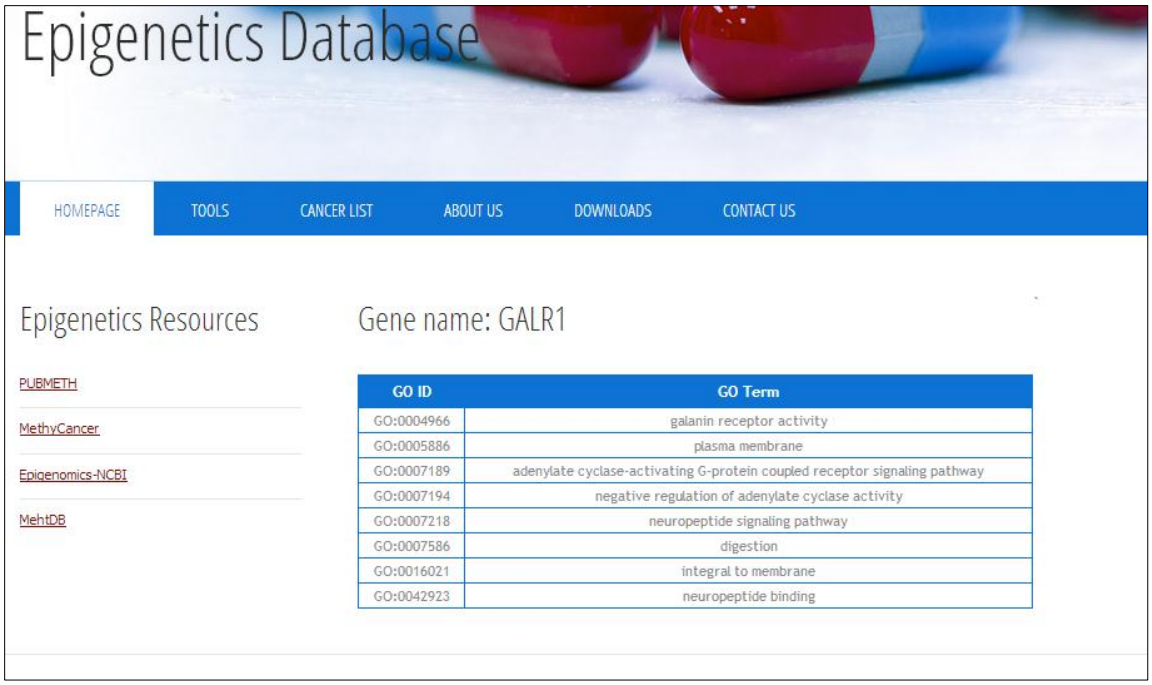


Figure 13 GO Terms associated with a methylated gene

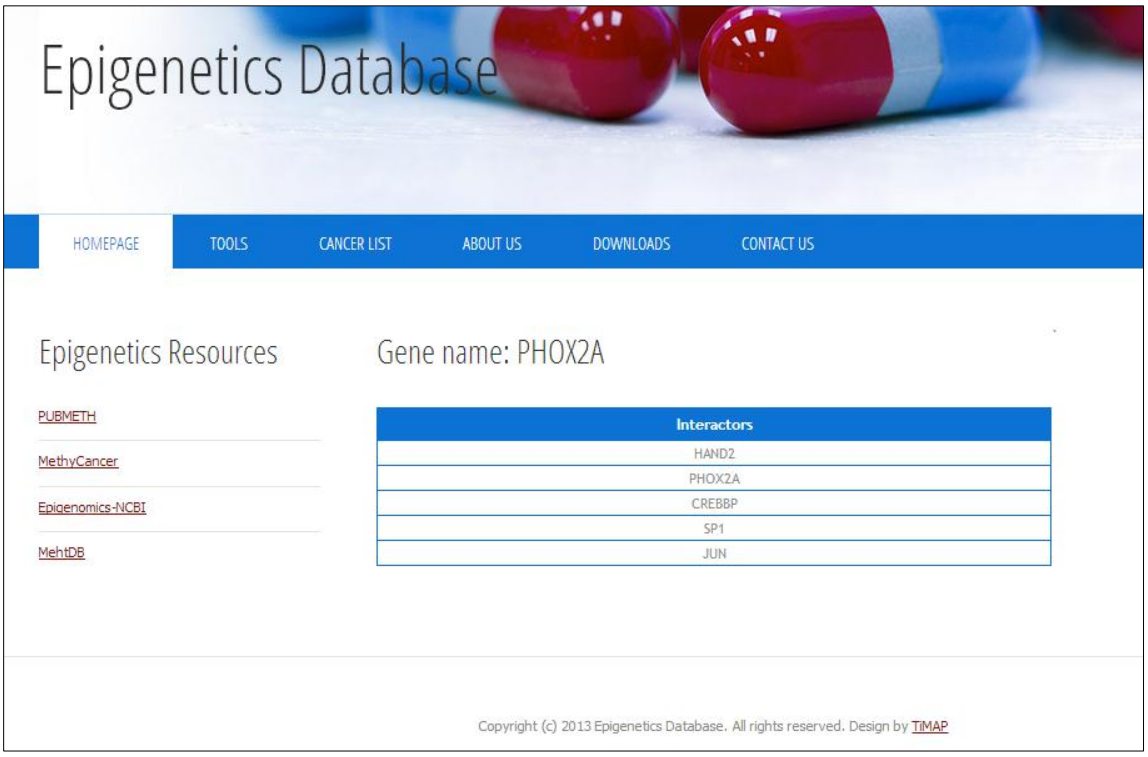


Figure 14 Interactions for a particular gene

Chapter 5 Conclusion

Our work proved the importance of integrating methylation data with the expression data for finding the patterns in a cancer. It showed the robustness of pipeline built for identifying key patterns which were helpful for distinguishing stages within a cancer and understanding critical targets for drug treatment. In addition to pipeline the data model used to store the meaningful data from analysis uses a sound parser, emphasizes the application of process automation and data warehousing for epigenetics.

The DNA methylation epigenetics data analysis pipeline is entirely based on the available TCGA data, which has the limitation of unequal samples, still we were able to prove the advantage of integrating epigenetic data, expression data and protein-protein interaction knowledge for advancing of systematic understanding of LUAD. This understanding can be further improved by incorporating the system biology approach to the epigenetic profile across the different stages of cancer data. This study's detailed analysis of epigenetics genes identified 72, 93 and 170 epigenetic genes across Stages I, II and III of LUAD. A set of 32 common epigenetic genes was identified across the three stages, and it was observed that methylation patterns were similar across Stages I and III, but were different in Stage II. The study also identified known and novel epigenetic genes across stages that were important in LUAD, these genes could be further validated in the laboratory for their scope as targets. The novel epigenetic genes identified were PTGDR, POU4F2, TLX3, and MMP26 along with these genes study identified early and late expression profiles of NEUROG1, AJAP1, and CORO6 in LUAD. System biology approach stated that epigenetic genes were not the hub nodes but could still affect the hub genes in the networks, eventually playing a critical role in the disease mechanism. Subnetworks of size 11 with 7 conserved genes across the three stages

were all literature validated, confirming their importance in LUAD. Therefore, it can be concluded that integrating epigenetic genes with expression data can be useful for comprehending in-depth disease mechanism and for the ultimate goal of better target identification.

The data model built to house the analysis results has a structure capable of managing the current input of data with flexibility to incorporate future changes which will come by integration of sequencing data. Currently the database holds analysis results for 3 cancers: Lung adenocarcinoma (LUAD), Colon adenocarcinoma (COAD) and Breast invasive carcinoma (BRCA). The web interface for this cancer results provides a structure and data which is more epigenetically meaningful compared to other databases which just display experimental data rather than compiling them in a way which makes sense to lot of scientist. Our database and its interface is designed in both simplistic as well as deep way that it will help a novice in the field to have it as its starting point and act as a support mechanism for a seasoned epigenetics analyst. The interface helps a user to query our database with multiple options which include a combination of 3 fields: cancer type, cancer stage and state of the genes. A user can select one of the cancer types that are currently available in our database and accordingly can select one of the stages or information on all stages. Once the cancer type and stage is selected an additional filter of methylation state is made available which helps them to extract genes whose state can be hypermethylated or hypomethylated. This provides the depth and granularity of analysis which is often required in understanding the epigenomics landscape of any cancer. On selecting above mentioned options a result table is displayed with multiple fields explaining the epigenetic parameters affected in that cancer. Beta-value field provides quantitative measure for a gene to determine its state in a cancer which can be either hyper or hypomethylated. The p-value field provides the

statistical significance measure which helps researchers to infer the qualitative meaning of beta-values. Additional fields such as Pathways, GO terms, Interactions and Disease help in understanding the relationship of gene with other cellular components covering most of the factors which can have influence on the activity of gene fulfilling the most important requirement of any epigenetic study. The disease field throws light on the role of gene in other diseases which brings in the additional information on the ways to tackle the gene aberration. Environmental factors field having the factors affecting the gene in a particular disease form help us to link these factors back to the state of gene in cancer. Thus our database provides a holistic view of DNA methylation epigenetic state of all affected genes in a particular cancer.

Hence using the pipeline developed through this thesis work and the data model designed to store the analysis results a researcher can create an experimental framework for any cancer study that he wishes to understand. The interface for database has a visual display which is created by a compilation of the fields that bring more epigenetic sense to the methylation data. Integration of methylation and expression data in the pipeline identifies important patterns and key targets which can help in prognosis of variable stages in cancer. The pipeline in conjunction with database can be utilized to design a strong architecture to detect the signs of aberrant DNA methylation of gene which can help in early detection of cancer.

Chapter 6 Discussion

Current epigenetic resources lack important parameters which make it difficult for researchers to picture a complete epigenomics working for any cellular machinery involved in cancer. There are numerous analysis pipelines available to analyze epigenetic data making it difficult to integrate them. We provide a solution to these problems by developing a universal pipeline not only to obtain significant methylated genes but to identify progressive methylation patterns present in different stages of any cancer. Although our pipeline was built using TCGA datasets, it can be extended to any platform providing beta-values and $\log_2$ ratios for the genes involved in cancer. The parser gives a great advantage in automating the execution of pipeline and saves the laborious tasks which often lead to human error.

Our pipeline for LUAD as a case study showed that the maximum number of DNA methylated genes was identified for Stage III followed by Stage II and then Stage I. None of the genes in Stage IV met the filtering criteria; therefore, no genes were identified as DNA methylated. From Table 1, it can be seen that hypermethylated genes were more prevalent in Stages I and III than in Stage II. Though this study identified 34 common DNA methylated genes (see Table 2) across the three stages, many of these have not been previously studied in LUAD. The HOX genes that were common across the three stages are TFs and grouped into four HOX families, A, B, C, and D; equivalent numbered HOX genes (HOXA9, HOXB9) in each family groups (A, B, C, D) are paralogs. The analysis found HOXA4, HOXA9, HOXB4, HOXD9, and HOXD12 genes with high methylation value, suggesting these genes play an important role in all stages of LUAD. These genes are known to be involved in cell proliferation while preventing apoptosis and help in survival [35], dysregulated behavior of HOX genes has been observed in ovarian cancer [36]. Early stage HOXA9 methylation has been identified in

lung cancer and used in early detection and prognosis [37, 38]. Our analysis found HOX genes in all stages, with hypermethylation in Stages I and III, hypomethylation in Stage II. While no previous studies have associated the profile of HOX genes with stages, though re-appearance was identified and our analysis demonstrated this aspect. Another gene identified by our analysis across all three stages was PTGDR, which was highly hypermethylated in Stages I and III (Table 1). PTGDR has been negatively correlated with smoking [39] and methylated in colon cancer [40], however, prior studies have not investigated its role in LUAD. POU4F2 and TLX3 were identified in all three stages and TLX3 was highly methylated in Stages I and III (Table 1). Previous studies have found them as methylated in leukemia and breast cancer respectively [41, 42] but not in LUAD. Overexpression of TBX20, which was also identified in this study (see Table 1) has been reported in lung cancer [43]. EVX1 and OTX2 (see Table 2) were identified as methylated in NSCLC and lung cancer [44, 45]. MMP26 has been associated with tumor development, invasion and metastasis of NSCLC but its methylation profile was not reported [46], our analysis showed it to be highly hypomethylated in Stage II (Table 1). There was no literature evidence about KPTAP8-1, REG2A, and SLX6 for their significance or methylation in lung cancer.

Of the 12 common methylated genes common to Stages I and II, LY96 has been previously associated with lung cancer [47]; ZNF577 and LVRN has been identified as methylated in lung cancer [45] and renal carcinoma, but not in LUAD [48]. LY96 was highly hypomethylated in Stage I and hypermethylated in Stage II (as shown in Table 1), suggesting further investigation into its role in LUAD.

In addition to pipeline the data model for database is only designed to capture the DNA methylation aspect of epigenetic study. In order to be called as epigenetic database the coverage needs to be extended to other wings of epigenetics which

include concepts such as histone modifications and data on chromatin remodeling. But the model provides a complete solution for DNA methylation data. It covers all important aspects that are needed for a researcher to understand the problem and provides sufficient fields to come with possible cure. The initial version of database was aimed to tackle the major epigenetic sub-type which is DNA methylation. Fields such as pathways, go terms, interactions are considered to be important factors affected by DNA methylation aberration. Understanding the relation between other cellular components can throw light on the role of affected gene and its cause of variation from normal state. The field for environmental factors included from KEGG database [24] includes the basic requirement for epigenetics analysis which most of the data resources in this sector fail to include in their data model. It laid a foundation which has given a way to a framework which can be used to include different data structures required to process different types of epigenetics data.

Thus the combination of analysis pipeline, parser implementing the pipeline and the data model housing the analysis results can be used for efficient processing of any DNA methylation epigenetic data.

Chapter 7 Future Work

Inclusion of other epigenetics branches such as histone modifications and chromatin remodeling data analysis forms the major task for the future work of our epigenetics system. We also wish to include sequence data which seems to be the future form for any kind of scientific data that would be available for analysis. For data model, it will be extended to accommodate these new forms of data that would be coming from the new pipelines designed. The interface will have more epigenetics analysis tools which researchers can use to analyze their datasets. The pipeline parser would also be made available through interface so that it can be used by researchers to automate their experimental analysis. We also wish to integrate various visualization tools such as Cytoscape [49] and Circos [50] through our interface which would help in visualization of analysis results.

Chapter 8 References

1. Wong AH, Gottesman II, Petronis A: **Phenotypic differences in genetically identical organisms: the epigenetic perspective**. *Human Molecular Genetics* 2005, **14**(suppl 1):R11-R18.
2. Bird A: **Perceptions of epigenetics**. *Nature* 2007, **447**(7143):396-398.
3. Russo VE, Martienssen RA, Riggs AD: **Epigenetic mechanisms of gene regulation**: Cold Spring Harbor Laboratory Press; 1996.
4. Jaenisch R, Bird A: **Epigenetic regulation of gene expression: how the genome integrates intrinsic and environmental signals**. *Nature genetics* 2003, **33**:245-254.
5. Straussman R, Nejman D, Roberts D, Steinfeld I, Blum B, Benvenisty N, Simon I, Yakhini Z, Cedar H: **Developmental programming of CpG island methylation profiles in the human genome**. *Nature structural & molecular biology* 2009, **16**(5):564-571.
6. Esteller M: **Cancer epigenomics: DNA methylomes and histone-modification maps**. *Nature Reviews Genetics* 2007, **8**(4):286-298.
7. Goelz SE, Vogelstein B, Feinberg A: **Hypomethylation of DNA from benign and malignant human colon neoplasms**. *Science* 1985, **228**(4696):187-190.
8. Schemies J, Uciechowska U, Sippl W, Jung M: **NAD+-dependent histone deacetylases (sirtuins) as novel therapeutic targets**. *Medicinal research reviews* 2010, **30**(6):861-889.
9. Plimack ER, Kantarjian HM, Issa J-P: **Decitabine and its role in the treatment of hematopoietic malignancies**. *Leukemia & lymphoma* 2007, **48**(8):1472-1481.
10. Lane AA, Chabner BA: **Histone deacetylase inhibitors in cancer therapy**. *Journal of Clinical Oncology* 2009, **27**(32):5459-5468.
11. Birney E, Stamatoyannopoulos JA, Dutta A, Guigó R, Gingeras TR, Margulies EH, Weng Z, Snyder M, Dermitzakis ET, Thurman RE: **Identification and analysis of functional elements in 1% of the human genome by the ENCODE pilot project**. *Nature* 2007, **447**(7146):799-816.
12. Ongenaert M, Van Neste L, De Meyer T, Menschaert G, Bekaert S, Van Criekinge W: **PubMeth: a cancer methylation database combining text-mining and expert annotation**. *Nucleic acids research* 2008, **36**(suppl 1):D842-D846.
13. He X, Chang S, Zhang J, Zhao Q, Xiang H, Kusonmano K, Yang L, Sun ZS, Yang H, Wang J: **MethyCancer: the database of human DNA methylation and cancer**. *Nucleic acids research* 2008, **36**(suppl 1):D836-D841.
14. Rollins RA, Haghighi F, Edwards JR, Das R, Zhang MQ, Ju J, Bestor TH: **Large-scale structure of genomic methylation patterns**. *Genome research* 2006, **16**(2):157-163.
15. Grunau C, Renault E, Rosenthal A, Roizes G: **MethDB—a public database for DNA methylation data**. *Nucleic acids research* 2001, **29**(1):270-274.
16. Fingerman IM, McDaniel L, Zhang X, Ratzat W, Hassan T, Jiang Z, Cohen RF, Schuler GD: **NCBI Epigenomics: a new public resource for exploring epigenomic data sets**. *Nucleic acids research* 2011, **39**(suppl 1):D908-D912.
17. Kitano H: **Systems biology: a brief overview**. *Science* 2002, **295**(5560):1662-1664.
18. Lokk K, Vooder T, Kolde R, Välk K, Võsa U, Roosipuu R, Milani L, Fischer K, Koltsina M, Urgard E: **Methylation markers of early-stage non-small cell lung cancer**. *PloS one* 2012, **7**(6):e39813.
19. Chen C, Grennan K, Badner J, Zhang D, Gershon E, Jin L, Liu C: **Removing batch effects in analysis of expression microarray data: an evaluation of six batch adjustment methods**. *PloS one* 2011, **6**(2):e17238.

20. Walter K, Holcomb T, Januario T, Du P, Evangelista M, Kartha N, Iniguez L, Soriano R, Huw L, Stern H: **DNA methylation profiling defines clinically relevant biological subsets of non–small cell lung cancer**. *Clinical Cancer Research* 2012, **18**(8):2360-2373.
21. Kruskal WH: **Historical notes on the Wilcoxon Unpaired Two-Sample Test**. *Journal of the American Statistical Association* 1957, **52**(279):356-360.
22. Storey JD: **A direct approach to false discovery rates**. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 2002, **64**(3):479-498.
23. Efron B, Tibshirani R: **An introduction to the bootstrap**, vol. 57: CRC press; 1993.
24. Kanehisa M, Goto S: **KEGG: kyoto encyclopedia of genes and genomes**. *Nucleic acids research* 2000, **28**(1):27-30.
25. Stark C, Breitkreutz B-J, Reguly T, Boucher L, Breitkreutz A, Tyers M: **BioGRID: a general repository for interaction datasets**. *Nucleic acids research* 2006, **34**(suppl 1):D535-D539.
26. Newman M: **Networks: an introduction**: Oxford University Press; 2009.
27. Watts DJ, Strogatz SH: **Collective dynamics of 'small-world'networks**. *Nature* 1998, **393**(6684):440-442.
28. Yu G, Li F, Qin Y, Bo X, Wu Y, Wang S: **GOSemSim: an R package for measuring semantic similarity among GO terms and gene products**. *Bioinformatics* 2010, **26**(7):976-978.
29. Chuang H-Y, Lee E, Liu Y-T, Lee D, Ideker T: **Network-based classification of breast cancer metastasis**. *Molecular systems biology* 2007, **3**(1).
30. Milenković T, Memišević V, Bonato A, Pržulj N: **Dominating biological networks**. *PloS one* 2011, **6**(8):e23016.
31. Vidal M, Cusick ME, Barabasi A-L: **Interactome networks and human disease**. *Cell* 2011, **144**(6):986-998.
32. Kroenke D, Auer DJ: **Database processing**: Science research associates Reading, MA; 1977.
33. Karolchik D, Baertsch R, Diekhans M, Furey TS, Hinrichs A, Lu Y, Roskin KM, Schwartz M, Sugnet CW, Thomas DJ: **The UCSC genome browser database**. *Nucleic acids research* 2003, **31**(1):51-54.
34. Sabbah C, Mazo G, Paccard C, Reyal F, Hupé P: **SMETHILLIUM: spatial normalization method for Illumina infinium HumanMethylation BeadChip**. *Bioinformatics* 2011, **27**(12):1693-1695.
35. Gray S, Pandha HS, Michael A, Middleton G, Morgan R: **HOX genes in pancreatic development and cancer**. *Jop* 2011, **12**(3):216-219.
36. Kelly ZL, Michael A, Butler-Manuel S, Pandha HS, Morgan R: **HOX genes in ovarian cancer**. *Journal of ovarian research* 2011, **4**(1):1-6.
37. Hwang S-H, Kim KU, Kim J-E, Kim H-H, Lee MK, Lee CH, Lee S-Y, Oh T, An S: **Detection of HOXA9 gene methylation in tumor tissues and induced sputum samples from primary lung cancer patients**. *Clinical Chemistry and Laboratory Medicine* 2011, **49**(4):699-704.
38. Rauch T, Wang Z, Zhang X, Zhong X, Wu X, Lau SK, Kernstine KH, Riggs AD, Pfeifer GP: **Homeobox gene methylation in lung cancer studied by genome-wide analysis with a microarray-based methylated CpG island recovery assay**. *Proceedings of the National Academy of Sciences* 2007, **104**(13):5527-5532.
39. Charlesworth JC, Curran JE, Johnson MP, Göring HH, Dyer TD, Diego VP, Kent JW, Mahaney MC, Almasy L, MacCluer JW: **Transcriptomic epidemiology of smoking: the effect of smoking on gene expression in lymphocytes**. *BMC medical genomics* 2010, **3**(1):29.

40. Spisák S, Kalmár A, Galamb O, Wichmann B, Sipos F, Péterfia B, Csabai I, Kovalszky I, Semsey S, Tulassay Z: **Genome-wide screening of genes regulated by DNA methylation in colon cancer development**. *PloS one* 2012, **7**(10):e46215.
41. Vilas–Zornoza A, Agirre X, Martin-Palanco V, Martín-Subero JI, San José-Eneriz E, Garate L, Alvarez S, Miranda E, Rodriguez-Otero P, Rifón J: **Frequent and simultaneous epigenetic inactivation of TP53 pathway genes in acute lymphoblastic leukemia**. *PloS one* 2011, **6**(2):e17012.
42. Hartmann O, Spyratos F, Harbeck N, Dietrich D, Fassbender A, Schmitt M, Eppenberger-Castori S, Vuaroqueaux V, Lerebours F, Welzel K: **DNA methylation markers predict outcome in node-positive, estrogen receptor-positive breast cancer with adjuvant anthracycline-based chemotherapy**. *Clinical Cancer Research* 2009, **15**(1):315-323.
43. Davis E, Teng H, Bilican B, Parker M, Liu B, Carriera S, Goding C, Prince S: **Ectopic Tbx2 expression results in polyploidy and cisplatin resistance**. *Oncogene* 2007, **27**(7):976-984.
44. Geng J, Sun J, Lin Q, Gu J, Zhao Y, Zhang H, Feng X, He Y, Wang W, Zhou X: **Methylation status of NEUROG2 and NID2 improves the diagnosis of stage I NSCLC**. *Oncology Letters* 2012, **3**(4):901-906.
45. Rauch TA, Wang Z, Wu X, Kernstine KH, Riggs AD, Pfeifer GP: **DNA methylation biomarkers for lung cancer**. *Tumor Biology* 2012, **33**(2):287-296.
46. Zhang Y, Zhao H, Wang Y, Lin Y, Tan Y, Fang X, Zheng L: **Non-small cell lung cancer invasion and metastasis promoted by MMP-26**. *Molecular Medicine Reports* 2011, **4**(6):1201-1209.
47. Rebhan M, Chalifa-Caspi V, Prilusky J, Lancet D: **GeneCards: integrating information about genes, proteins and diseases**. *Trends in Genetics* 1997, **13**(4):163.
48. López-Lago MA, Thodima VJ, Guttapalli A, Chan T, Heguy A, Molina AM, Reuter VE, Motzer RJ, Chaganti RS: **Genomic deregulation during metastasis of renal cell carcinoma implements a myofibroblast-like program of gene expression**. *Cancer research* 2010, **70**(23):9682-9692.
49. 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.
50. Krzywinski M, Schein J, Birol İ, Connors J, Gascoyne R, Horsman D, Jones SJ, Marra MA: **Circos: an information aesthetic for comparative genomics**. *Genome research* 2009, **19**(9):1639-1645.

Chapter 9 Appendix

9.1 Pathway Distribution across stages for LUAD

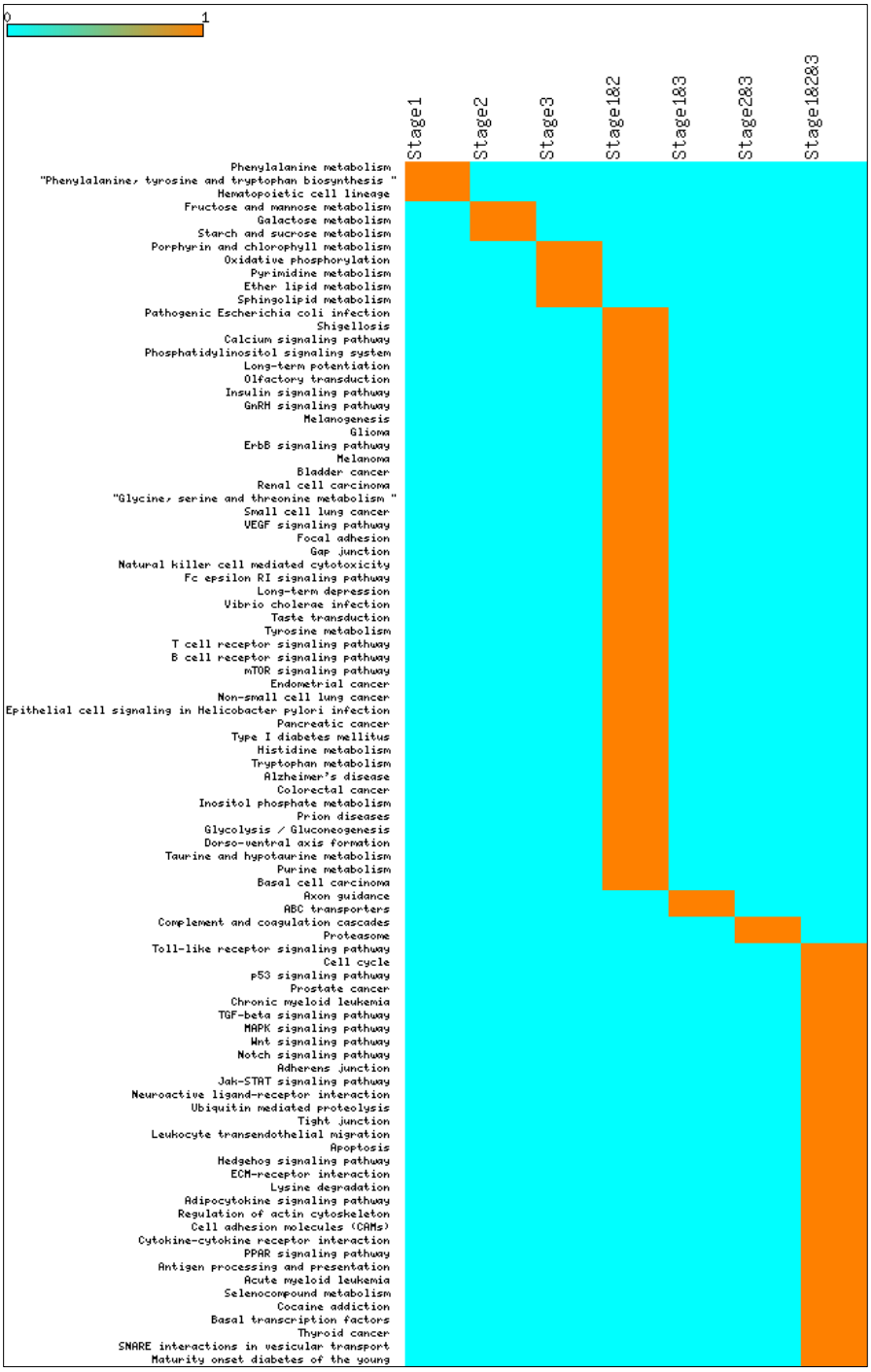


Figure 15 Pathway Distribution

Table 9 Pathway distribution according to sub-network sizes

Stage	Sub-network Size	Pathway distribution			
		Cancer	Lung cancer	Signaling	Metabolic +others
I	2	1	-	3	5
	3	4	1	5	10
	4	68	45	175	568
	5	2685	1466	5072	1263
II	2	3	1	2	6
	3	58	19	69	107
	4	1176	532	2049	4230
	5	31982	15884	59137	133380
III	2	3	1	3	7
	3	80	33	95	155
	4	1476	679	2578	5952
	5	141149	19951	76205	175691

Table 10 Analysis of common and unique sub-networks of size 4 revealing the significant genes

Pathway	Stages						
	I & II & III	I & II	I & III	II & III	I	II	III
Cancer	18	4	43	820	-	336	591
Lung cancer	11	-	25	369	-	153	274
Signaling	70	2	222	1372	27	641	1049
Metabolic + others	74	-	792	2844	135	1347	2677

Table 11 DNA methylated genes in UBC subnetworks across stages

Stage	UBC interaction with methylated genes
I	FOXG1, GAS7, HLA-G, HOXD8, LY96, MSC, NPY, PHOX2A
II	ACTN2, CDO1, FOXG1, HOXA1, HOXD4, HTR2C, INPP5B, LHX1, NEFH, OTX2, PHOX2A, PRKCB, SERPINB5, SLC6A2, SRGN, and TAL1
III	ATP6VOD2, CFTR, CRMP1, DGKI, EPO, FLG, HAND2, HOXA7, HOXB4, HOXD4, HOXD12, IHH, INPP5B, NECAB2, NEUROG1, NPY, PDZRN3, POU3F1, PHOX2A, SGMS2, SLC6A2, SPTA1, TBX5, TMEM132D, WNK2 and XDH

9.2 P-value correction-before and after resampling and bootstrapping

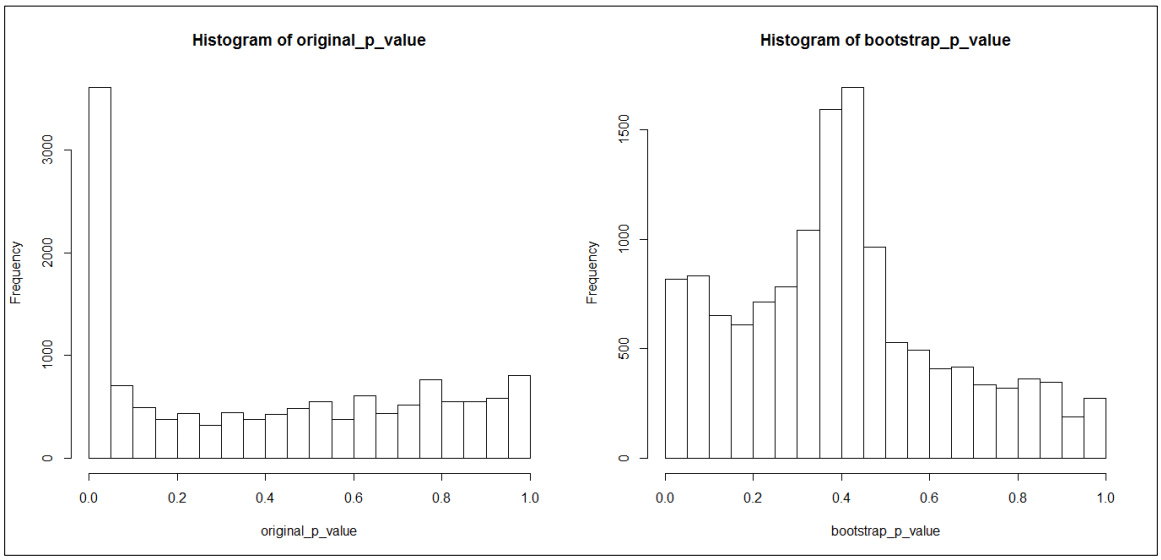


Figure 16 Resampling p-value correction

9.3 Epigenetic Data Annotation

Table 12 Functional annotation

Database	Fields in Database	Reason for selecting DB
BioGRID	• interactor_A • interactor_B • synonyms_interactor_A • synonyms_interactor_B • experiment_type • score • source_database	BioGRID database provides information on physical interactions for methylated genes which can be used to understand the network of genes in a particular cancer.
GO database (gene2go)	• entrezid • go_id • go_term	GO database provided the functional annotations to the gene.
KEGG (KEGG_disease)	• Entry • Name • Description • Category • Pathway • Gene • Env_factor	Disease table from the KEGG database is used to include information on disease caused due to malfunctioning of pathways and environmental factors responsible for disease.
KEGG (KEGG_gene_pathway)	• name • pathway_id • entrezid	Gene-pathway table from KEGG helps to extract pathways affected by the gene
Entrez (Entrez2gene)	• entrezid • gene	Entrez2gene table is used for mapping entrezid to human readable gene symbols.
UCSC genome browser (UCSC2GeneSymbol)	• kgid • genesymbol	UCSC2GeneSymbol table is used for mapping keg gene ids to human readable gene symbols.

AKSHAY DESAI

532 Drake Street, Indianapolis, IN-46202. Ph. 317-478-1666, akdesai@iupui.edu

Qualification Summary

- Efficient at the computational and analytical skill sets developed over the years in bioinformatics research and internship at Dow AgroSciences.
- Proficient in various web technologies and scripting languages. Build "next generation sequencing pipeline" and "biological analytical tools" which are used at research institutes for processing high throughput data.
- A highly motivated team player with flexible thinking that is enthusiastic with working in diverse roles.

PROFESSIONAL EXPERIENCE:

INTERN, Dow AgroSciences.
May 2013-September 2013

1. Worked with ITDA (Information Technology and Data Analysis) department supporting biological discovery function on project testing multiple systems integration ensuring accuracy of scientific data flow.
2. To ensure appropriateness of testing interacted closely with "developers and system owners" to formulate test cases and test scripts for black box testing and integration of the systems.
3. In an effort to identify issues executed test scripts by automating them to uncover defects early enough translating to saving over significant man hours of time and ensuring more stable delivery of the system.

RESEARCH EXPERIENCE:

Biomedical Text Mining Group, Indiana University

Graduate research assistant
Sept 2011 - Present

1. **Thesis:** "Finding the epigenetic modular progression across different stages of LUAD."
2. Characterization of the Korean Genome.

CLASS PROJECTS:

1. Developed a Dental Information Management System (DIMS) which helped dental professionals to manage information regarding patients and their business.
2. Build a rank based database for methylated genes in cancer using publically available expression datasets helping researchers to understand epigenetics with statistical measures.
3. Constructed an integrated web social health network to facilitate sharing and awareness of health institutes. (**Engineering and Technology Department**)
4. Performed an integrative whole genome analysis of TCGA breast cancer sequencing data to identify significantly mutated noncoding RNAs using NGS tools. (**IU School of Medicine**)

EDUCATIONAL QUALIFICATIONS:

Master of Science, Bioinformatics (Currently Pursuing)

Expected graduation date: October 2013

Indiana University, School of Informatics, Indianapolis, IN Current GPA: 3.69/4

Advanced Diploma, Bioinformatics
May 2010

Rajiv Gandhi Institute of IT & Biotechnology, Bharti Vidhyapeeth University, Pune, India

GPA: 3.6/4

Bachelor of Science, Biotechnology
May 2009

Modern College, Pune University, Pune, India GPA: 3.5/4

COMPUTATIONAL QUALIFICATIONS:

Languages: Python, Perl, JAVA, C, Base SAS, R, SPSS, Shell scripting

Web Technologies & Frameworks: HTML, PHP, Perl CGI, Python CGI, Java scripts

Packages: R and BioConductor

Database and Development tools: Eclipse, Net Beans, EasyPHP, Aqua Data Studio.

Database: MySQL, Oracle 10g: SQL, PL/SQL

Operating System: Windows XP, Vista and 7, UNIX

BIONFORMATICS QUALIFICATIONS:

Microarray analysis: GenePattern, GSEA, MeV

Visualization tools: Cytoscape, IGV

NGS analysis: SAM tools, BAM tools, VCF tools, GATK, ANNOVAR, CNV-seq tool, snpEFF.

HONORS:

1. Won best student of the year award in Advanced Diploma course.
2. Mentored and volunteered for project seed under CTSI, Indianapolis.
3. Won 2nd prize at “CAMDA 2013” conference held in Berlin for Big Data Analytics. (http://dokuwiki.bioinf.jku.at/doku.php).