Analysis of DNA methylation and RNA-Seq data for Prostate Adenocarcinoma: An Integrative Approach

Arpit Singh

Delhi Technological University, Delhi, India

ABSTRACT

Epigenetics is rapidly gaining recognition as it accounts for the change in phenotype without any change in the genotype or the DNA sequence. These are the heritable changes observed in gene expression without any change observed in the coding sequence. DNA methylation is the most extensively studied epigenetic mechanism which adds a methyl group to DNA at a cytosine base almost every time accompanied by a guanine base. These sites of methylation are also called “CpG” islands known to harbour promoter regions. Hence methylation at promoter regions directly affects the binding of DNA-binding protein thus inhibiting transcription and gene expression.

DNA methylation is observed to play major roles in gene regulating mechanisms and gene silencing mechanisms. Therefore understanding the relationship between DNA methylation and gene expression becomes very important. As a proof of concept, Prostate Adenocarcinoma (PRAD) was chosen as the cancer to be studied as it is the second leading cause of death in men. DNA methylation (level 1) and Gene expression data (level 3) from The Cancer Genome Atlas (TCGA) were downloaded for 18 normal matched with tumor and 18 tumor matched with normal samples from batch 184.

“R” programming language was used to integrate and analyse the data. “R – Bioconductor” packages “minfi” and “COHCAP” were used to find 453 differentially methylated regions with with p-value < 0.05, fdr < 0.05 and beta value (methylation) > 0.2.

The gene expression data was integrated with matched TCGA IDs and Pearson correlation analysis was carried out. 180 significant correlations were identified, out of which 112 correlations were chosen by applying stringent rules like correlation < -0.5, p value < 0.001 and false discovery rate < 0.001. Upon visual inspection of the results, 74 correlations were finally filtered and functional enrichment was carried out. It was discovered that genes "GSTP1" and "FGFR2" are already known to be involved in prostate cancer pathway and progression and these genes were present in the final filtered significant correlations. This approach may indicate the involvement of other novel genes in the prostate cancer pathway for which experimental validation must be carried out.

INTRODUCTION

The classic genetic model alone cannot account for the variation in phenotype of the entire population (Esteller, 2008). It is evident that there are factors (other than the genetic make-up) involved in the phenotypic variation observed in a population. Epigenetic inheritance is one such mechanism widely studied across the globe which accounts for the "other" factors involved in the phenotypic variation. This represents cellular information other than the mere sequence of the DNA (Feinberg and Tycko, 2004).

DNA methylation is the most common form of epigenetic mechanism studied by research groups all over the world (Egger *et al*., 1999). It involves addition of a methyl group to C-5 position of cytosine of DNA with the aid of methyl transferases (Feinberg and Tycko, 2004) and thus make it unavailable to bind to DNA-binding proteins (Jones and Takai, 2001). DNA methylation most frequently occurs at cytosine immediately followed by a guanine molecule also called a CpG island (Feinberg and Tycko, 2004). It is known to play an important role in regulation of gene expression (Jones and Laird, 1999; Esteller, 2008).

The Cancer Genome Atlas (TCGA) is a common platform designed to distribute and handle large volumes of research data for more than 20 types of cancer (*https://wiki.nci.nih.gov/display/TCGA/About+TCGA*). Prostate Adenocarcinoma (PRAD) is a common type of cancer prevalent in western countries and is a major cause of death in men (Hsing and Chokkalingam, 2006). TCGA provides information about the methylation changes observed in the tumor and normal samples along with gene expression information and somatic changes and variations observed. The smart architecture of TCGA enables a researcher to download raw or processed data wherever applicable and available. Independent studies can be carried out to compare, analyse and interpret the information from various platforms on a particular sample. As is the mission of TCGA, the atlas of changes can be analysed and stored to reduce the gap in the cancer and its molecular biology.

Data analysis involves handling large data using suitable programming language like "R" and employing the use of several packages designed to carry out the analysis. In this study, DNA methylation data (level 1) and Gene expression RNA-Seq v2 data (level 3) was downloaded for PRAD of batch 184 with 18 normal with matched tumor and 18 tumor with matched normal samples. Statistical analysis was carried to identify the regions differentially methylated amongst the normal and tumor samples and integrate the information with RNA-Seq v2 data to find out the significantly correlated genes where a relation between methylation and gene expression can be observed.

In this study, "R – Bioconductor" package COHCAP was used to identify the differentially methylated regions and gene expression data was integrated using the matched TCGA IDs Data pre-processing was carried out using tools like MS-Excel, "R" programming language and "R – Bioconductor" package minfi. The results were visualized in the form of box and scatter plots and graphs and viewed on Integrative Genomics Viewer (IGV, Broad institute) (Robinson *et al*., 2011; Thorvaldsdóttir *et al*., 2013).

REVIEW OF LITERATURE

3.1 Epigenetics

The study of genetics simply aims at studying the genotype responsible for a quantifiable phenotype. It points to the relation of a gene with the proteins expressed and further causing normal or abnormal body function. Any kind of genetic change is usually irreversible and stable. Epigenetics refers to the changes in the expression of the genotype which are heritable but there are no changes in the contributing DNA sequence causing a change in the phenotype (Bird, 2007). A good example could be in developmental biology, all cells contain same DNA but differentiate to form different cells and further different tissues and in turn have different body functions and roles (Jones *et al*., 1999).

Epigenetic changes in the body are natural occurring instances and regularly occur to carry out normal bodily functions and routine (Egger *et al*., 2004). These changes could be readily influenced by external factors like environmental exposures, age, lifestyle/trends and the state of disease.

Epigenetic inheritance which are currently known are (Feinberg *et al*., 2004):

1. DNA Methylation
2. Histone modification
3. Genomic imprinting
4. Non-coding RNA regulatory system (Riddihough *et al.*, 2010)

The most widely studied epigenetic change is DNA Methylation and still it is not clearly understood by the researchers (Jones *et al*., 2001). DNA Methylation negatively correlates with the gene expression.

Epigenetic changes are reversible and play an important role in regulating many cellular processes (Simmons *et al.*, 2008). Without changing the DNA sequence, these epigenetic changes bring out a quantifiable phenotypic change. Any disturbance in the epigenetic balance can cause several maladies like cancer, neurodegenerative disorders and chromosomal disorders (Egger *et al*., 2001). The eagerness to learn and demystify the mystery of epigenetics has led to growth in research in epigenetics.

3.2 DNA Methylation

DNA Methylation is the most widely studied epigenetic mechanism (Egger *et al*., 1999). Events have been reported of DNA Methylation portraying gene silencing behaviour (Jones *et al*., 1999). DNA Methylation has been reported to show negative correlation with gene expression as it inhibits initiation of transcription (Holliday, 2006).

The mechanism of DNA Methylation involves adding a methyl group to DNA. This chemical process is very specific and is usually observed in a CpG region where a cytosine (C) nucleotide and guanine (G) nucleotide are next to each other linked by a phosphate group (Egger *et al.*, 2004; Jones & Baylin, 2002; Robertson, 2002; Feinberg *et al*., 2004). Methyl

group addition is brought about by a family of enzymes namely DNA methyltransferases (DNMTs). DNMT1, DNMT3a and DNMT3b play important roles in maintenance of DNA methylation routine (Baylin, 2005).

These CpG sites are usually found in 5' end of the gene regulatory architecture which adds to the explanation of inhibiting transcription of genes (Esteller, 2008). These CpG sites can be on a single Cytosine base or multiple. Cpg islands are rich in CpG sites but the exact criteria of a CpG island is not well defined or universally accepted. A lot of investigation goes in analysing the role of DNA methylation at these CpG islands rather than elsewhere as these CpG islands are observed to have strong promoter activities (Jones *et al*., 2001).

Just as important as DNA methylation is DNA demethylation as the DNA methylation is reversible and hence DNA demethylation plays an important role in genes reprogramming (Morgan *et al*., 2005). This fact paves way for the advent of various epigenetic therapies and drug design (Egger *et al*., 2004).

3.3 DNA Methylation arrays

Although the advent of Next-Generation sequencing has out-shadowed the classic microarray techniques, microarray is still strong and living attributing to its cost effectiveness, robust optimization over the years and rich experience (Meaburn and Schulz, 2012). Small laboratories cannot afford the very high costs of Next-Generation sequencing and hence trade off its high precision accuracy with low cost microarray techniques.

Microarrays are designed where most of the CpG sites all across the genome are interrogated. A myriad of probes are designed for CpG sites covering all parts of the human genome. This design is consulted by a panel of reviewers consisting of researchers and scientists (*Illumina HumanMethylation450K*). The main principle behind DNA methylation microarrays is bisulphite conversion step. Upon Bisulphite Conversion and whole genome amplification, the analysis of DNA methylation is reduced to an analysis of single nucleotide polymorphisms (SNPs) for T's and C's. If C is found, original cytosine was methylated and if T is found, original cytosine was non-methylated. The flowchart below depicts the basis behind DNA methylation microarray.

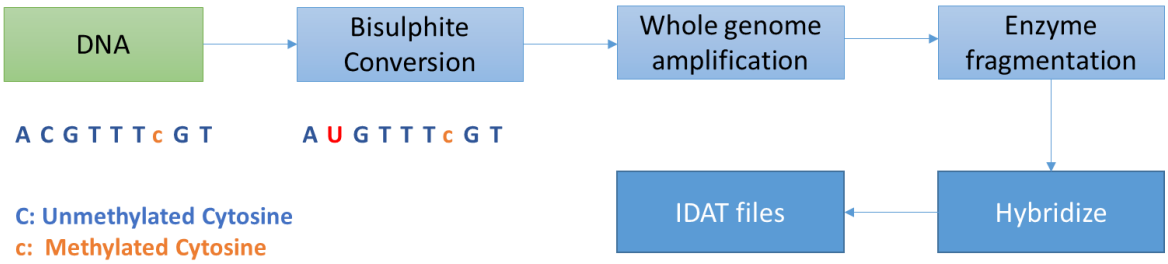

Figure 1: DNA methylation microarray

3.4 RNA-Seq: Gene expression profiling

RNA-Seq (RNA Sequencing) is a next-generation sequencing technique that captures the transcriptome of the organism (Chu and Corey, 2012). RNA-Seq boasts of high accuracy and precision owing to the next-generation sequencing abilities. It reveals the entire transcriptome of the organism at that moment which refers to the transcripts and their isoforms. In comparison to the existing gene expression microarrays, RNA-Seq data clearly stands out because of accurate results, better resolution and it doesn't require the need of previously known genomic sequences of the organism or a model organism (Wang *et al*., 2009).

3.5 Platforms

The platforms used for the data used in this study are depicted in the table below (TCGA wiki: *https://wiki.nci.nih.gov/*):

Platform Code	Platform Alias	Platform Name
HumanMethylation450	HumanMethylation450	Illumina Infinium Human DNA Methylation 450
IlluminaHiSeq_RNASeqV2	IlluminaHiSeq_RNASeqV2	Illumina HiSeq 2000 RNA Sequencing Version 2 analysis

Table 1: TCGA platform

3.6 Data Matrix (TCGA)

The Cancer Genome Atlas (TCGA) is a common data storage portal that provides data generated by the initiatives taken by TCGA and the collaborative research laboratories. The data comprises of clinical information, genomic information, variants information, SNP information, DNA methylation information and gene expression information of the tumor and normal genomes. TCGA contains information about the tumor samples mapped with the normal samples and normal samples mapped with the tumor samples. TCGA makes use of the latest technology and platforms. TCGA allows the users to freely use and integrate data downloaded from the TCGA portal. If the data is not made available it cannot be downloaded. This study uses the freely available that is downloaded from the TCGA Data Matrix portal (URL: *https://tcga-data.nci.nih.gov/tcga/dataAccessMatrix.htm*).

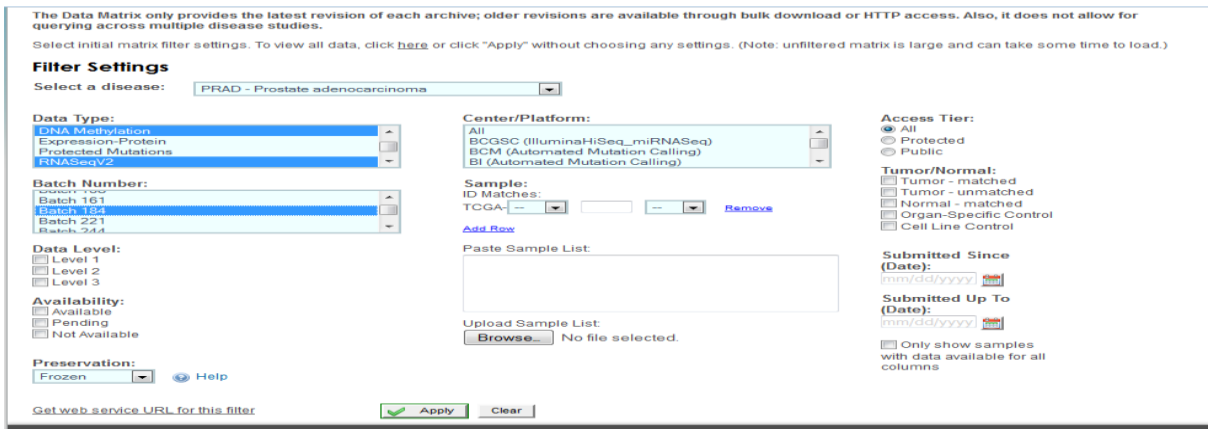

Figure 2: TCGA Data Matrix

In order to use the TCGA data one must be well versed with the guidelines, bar-codes and the general architecture of the TCGA (*https://wiki.nci.nih.gov/display/TCGA/TCGA+barcode*). A wiki page of TCGA is available containing all the information a researcher will require when analysing the off-shelf data.

To analyse the DNA Methylation data, TCGA barcode is very important to be understood in this case. One must also understand the level of data he/she must deal with (https://tcga-data.nci.nih.gov/tcga/tcgaDataType.jsp). In this study level 1 type data for DNA Methylation is retrieved and for RNA-Seq data, level 3 type data was used.

Data Type Name	Level	Description
RNASeqV2	3	The calculated expression signal of a gene, per sample File type: tab-delimited (.txt)
Array-based DNA Methylation	1	Raw signal intensities of probes for each participant's tumor sample File type: tab-delimited (.txt) and binary (.idat)

Table 2: TCGA Data Level

TCGA barcode is well documented code. The identifier for the sample type is the fourth identifier after the hyphen. Tumor types range from 01 – 09, normal types from 10 – 19 and control samples from 20 – 29 (URL: *https://wiki.nci.nih.gov/display/TCGA/TCGA+barcode*). For examples in "TCGA-02-0001-**01C**-01D-0182-01", identifier highlighted represents that this is a tumor sample.

3.7 Statistical analysis of DNA Methylation data

The data downloaded from the TCGA Data Matrix portal is large in size and therefore "R" programming language was used to analyse the data. R studio was used to run the R session for better graphical interface and better operation. Bioconductor package is open source R based software suite that provides tools for analysis of high-throughput biological data (Gentleman *et al*., 2004).

To analyse DNA methylation data, several R packages are in use by the scientific community. In this study we use the Bioconductor packages "COHCAP" and "minfi".

"minfi" is a Bioconductor package that is used to read the raw IDAT files generated from Illumina 450k array. These raw files are pre-processed and methylation values are extracted for further use in "COHCAP" pipeline (Aryee *et al*., 2014).

City of Hope CpG Island Analysis Pipeline (COHCAP), is a workflow developed to analyse the DNA methylation data produced by microarray or sequencing. The pipeline identifies differentially methylated regions (DMRs) by identifying CpG islands where CpG sites show a consistent pattern of methylation. By default the minimum number of CpG sites in an island is 4 which can be modified to obtain densely populated islands.

COHCAP is the only R based algorithm that integrates the methylation data with normalized expression data. COHCAP detects significant negative correlations that are most likely involved in the regulation of the gene exression where the CpG island lies. It generates

quality control graphs to filter out technical defaulters. Excel sheets or text files are created for the user to further apply stringent rules to the data. For visualization, wiggle files are generated that can be viewed in Integrated Genomics Viewer (Robinson *et al*., 2011) or UCSC Genome Browser (Warden *et al*., 2013).

3.8 Prostate Adenocarcinoma

Prostate Adenocarcinoma is a common type of cancer that accounts for almost all prostate cancers (*http://www.cancer.gov/types/prostate*). It is characterized by the development of cancer in the prostate gland. It is more prevalent in the western countries and is the second leading cause of deaths in men (Hsing and Chokkalingam, 2006). A lot of research efforts are put in to identify the aberrant genes and identify a molecular model for Prostate Adenocarcinoma (Porkka and Visakorpi, 2004).

The genes currently identified as those involved in the development of Prostate Cancer are GSTP1, PTEN, TP53, and AR (Porkka and Visakorpi, 2004). From the KEGG disease database, following gene entries were obtained (*http://www.kegg.jp/dbget-bin/www_bget?ds:H00024*).

Genes (KEGG Disease)	
AR	(amplification, mutation)
CDKN1B	(allelic loss)
NKX3.1	(allelic loss)
PTEN	(allelic loss)
GSTP1	(hypermethylation)
TMPRSS2-ERG	(translocation)
TMPRSS2-ETV1	(translocation)
TMPRSS2-ETV4	(translocation)
TMPRSS2-ETV5	(translocaiton)
SLC45A3-ETV1	(translocation)
SLC45A3-ELK4	(translocation)
DDX5-ETV4	(translocaiton)

Table 3: Genes (KEGG Disease)

Another gene "FGFR1" is also essential for the Prostate cancer development and progression (Yang e al., 2013). EZH2 and KAI1 are also involved in the Prostate cancer (Dong *et al*., 1995; Varambally *et al*., 2002).

3.9 Objectives

The inspiration to carry out this study comes from the fact that due to the advent of next generation sequencing and high performance microarrays, a lot of data is generated from various research laboratories and as a bioinformatician, one must work to apply his/her skills to make sense of the data. Further innovative methodologies can be used to analyse data in an intelligent way. In this research, following objectives were fulfilled:

1. Download matched data for DNA Methylation and RNA-Seq for 36 samples (18 Normal + 18 Tumor) and preprocess the data using R programming language.

2. To find differentially methylated regions and integrate them to the expression data (normalized level 3 data).
3. To visually inspect and analyze the correlation between methylation and expression using a genome browser.

3.10 Pipeline used in current study

In this study the following pipeline/workflow is used:

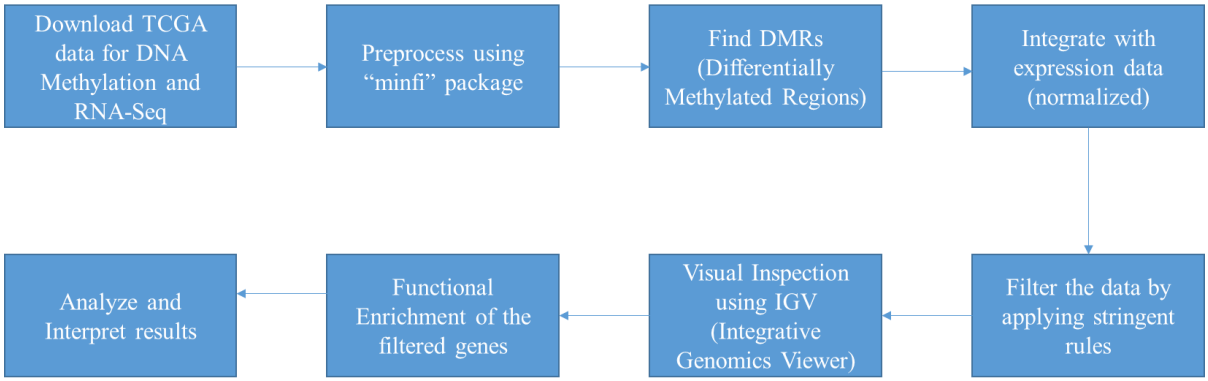

Figure 3: Workflow

METHODOLOGY

4.1 Data retrieval

The 450k Illumina methylation array data and RNA-Seq v2 data for a total of 36 samples was collected from TCGA (The Cancer Genome Atlas) using the data matrix portal by providing the TCGA IDs for matched tumor and normal samples for batch 184 and disease "Prostate Adenocarcinoma (PRAD)". There were 18 normal matched with tumor samples and 18 tumor matched with normal samples.

Data Type	**DNA Methylation**	**RNA-Seq**
Level	1	3
Center/Platform	JHU_USC (HumanMethylation450)	UNC (IlluminaHiSeq_RNASeqV2)
Batch	184	184
Disease	PRAD	PRAD

Table 4: Data description

4.2 Data pre-processing

R programming language was used to pre-process the data. Since the data is large, "R" was used to handle it. For the JHU_USC (HumanMethylation450) data, a sample description file is to be made to be able to feed it in another R package "COHCAP" for further analysis. Therefore using in-house scripts the sample description file was generated for 36 samples. For UNC (IlluminaHiSeq_RNASeqV2) data, an integrated expression value file was to be created for all the matched samples. In-house scripts were used to extract the normalized expression values from the level 3 data and was integrated in a single file for further analysis.

For the expression values, a log2 transformation was carried for better visualization and comparison. Since a lot of samples contained zero values, log2(x + 1) was carried out.

4.3 Data analysis

DNA Methylation and RNA-Seq data was then analysed using R package "COHCAP". Differentially methylated regions were analysed for the methylation data and the filtered islands were mapped to the corresponding genes and this was integrated with normalized expression values for each matched sample.

4.4 Data visualization

The results obtained by COHCAP analysis were viewed using Nitro Reader 3, Microsoft Excel and the .wig files generated by differentially methylated region analysis were viewed and interpreted in IGV (Integrative Genomics Viewer), Broad Institute (Robinson *et al*., 2011; Thorvaldsdóttir *et al*., 2013).

4.5 Software and Hardware

A Linux (Ubuntu) based system with 6 GB RAM was used to pre-process and analyse the data. "R" was installed on Ubuntu and Bioconductor packages were used for the pipeline. R studio was used for a better graphical user interface and easy operation.

Package/Tool	Description
Minfi	Analysis of 450K array data
COHCAP	Pipeline for CpG island analysis of 450K array data
Integrative Genomics Viewer, Broad Institute	Visualize the .wig files to inspect the methylation changes and their corresponding position with respect to the reference human genome.

Table 5: Softwares and tools

4.6 Functional enrichment and Pathway Analysis

Pathway analysis for the filtered genes was carried out by the Kyoto Encyclopaedia of Genes and Genomes (KEGG) database and the Database for Annotation, Visualization and Integrated Discovery (DAVID) database (Huang *et al*., 2009).

4.7 "R" scripts

R scripts developed locally were used to pre-process the DNA methylation data and RNA-Seq data. The scripts are included in the appendix files attached at the end.

4.8 Quality Control

Some of the sample data were outliers and that affected the differential methylation analysis and integration with expression data analysis. Such findings were removed to avoid false positives due to unexplainable abnormal or zero values that can be attributed to experimental artifacts or technical problems. The filtered genes were visualized in Integrated Genomics Viewer and their correlation plots were inspected. The genes where the correlation or the clustering of normal and tumor samples was not clear were removed.

RESULTS

5.1 Clustering of the Tumor and Normal samples

The COHCAP pipeline involves a quality control step where the samples are clustered and plotted for the user to remove unwanted samples or outliers.

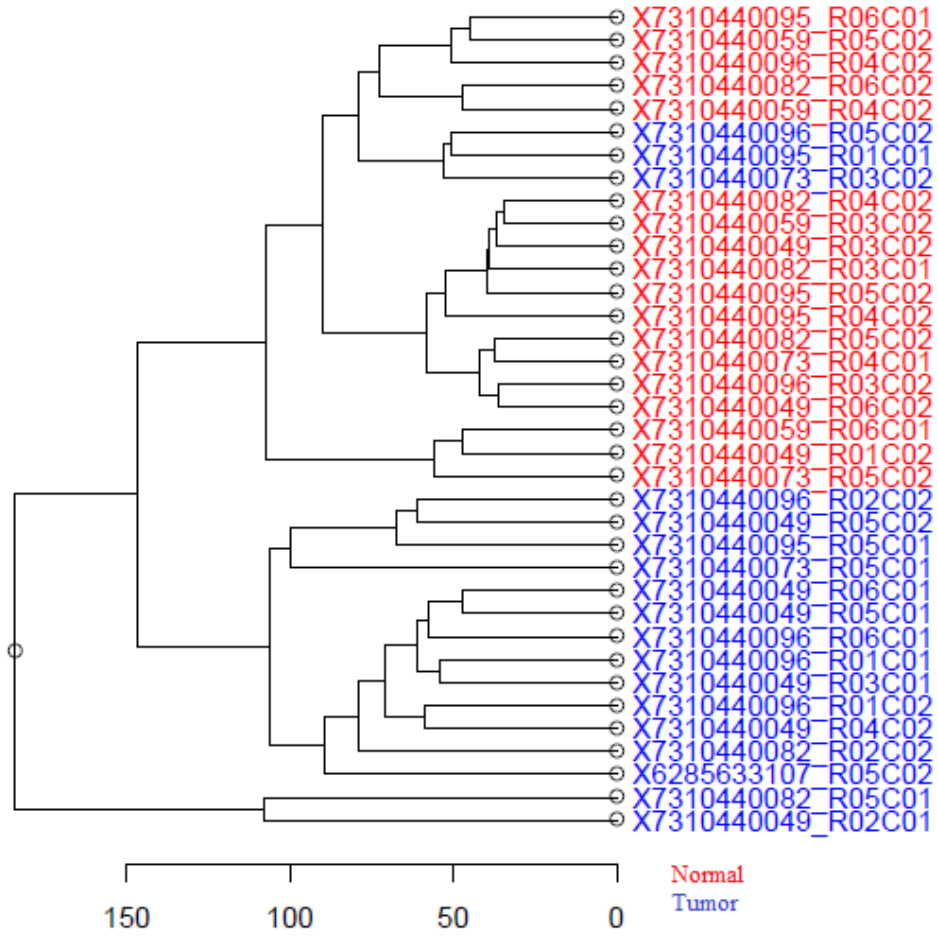

Figure 4: Sample Clustering

The normal and tumor samples were clustered within their own groups as it can clearly be seen in the above plot where normal samples are represented by red color and tumor samples are represented by blue color.

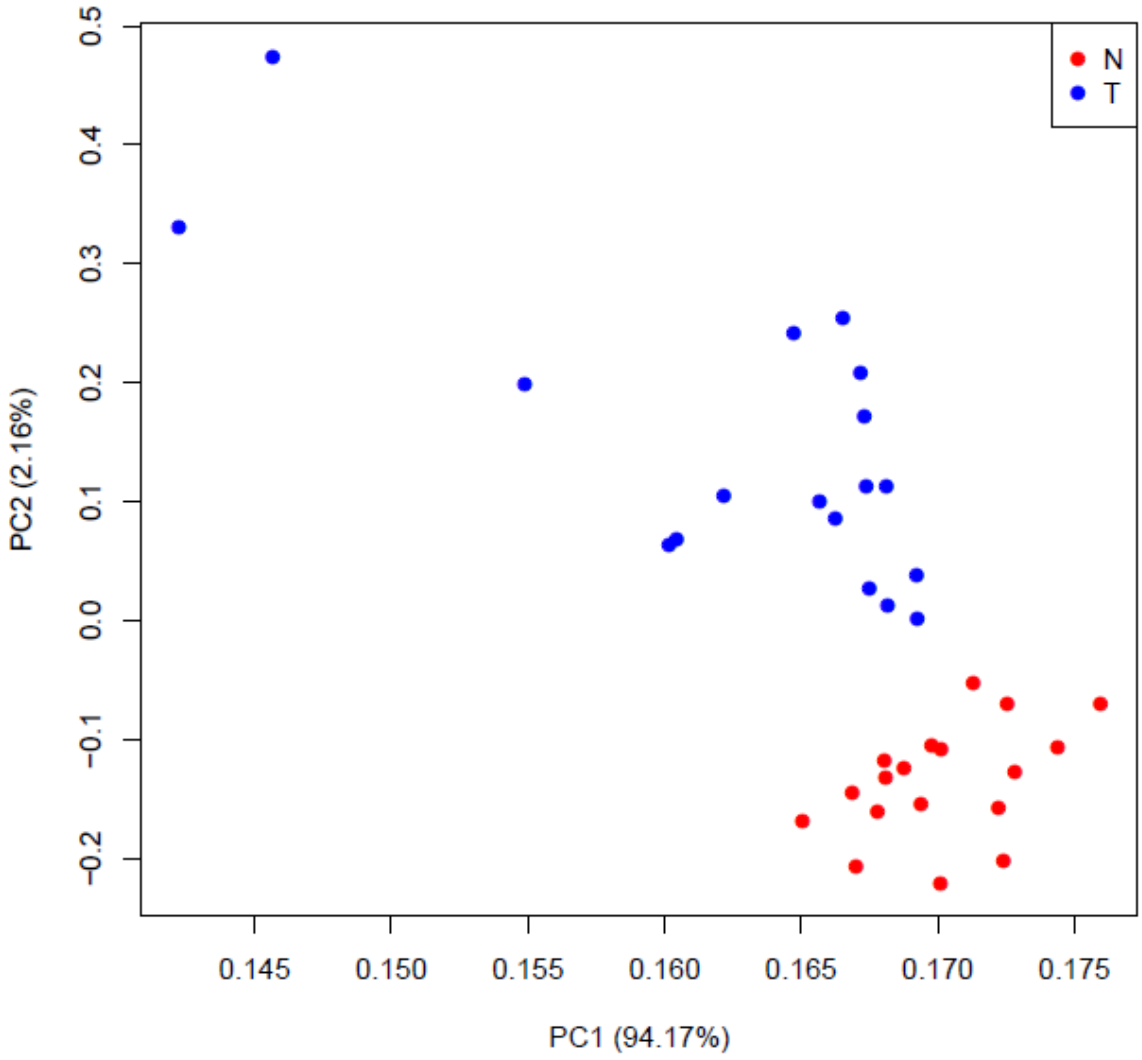

Figure 5: PCA

5.2 Quality Control

The histogram plots of the methylation values (beta values) of the sample were plotted to check if there is any sample that is a product of technical problem or an experimental artefact. The plot is shown below and it can be seen that all the samples follow normal methylation distribution usually seen in these experiments.

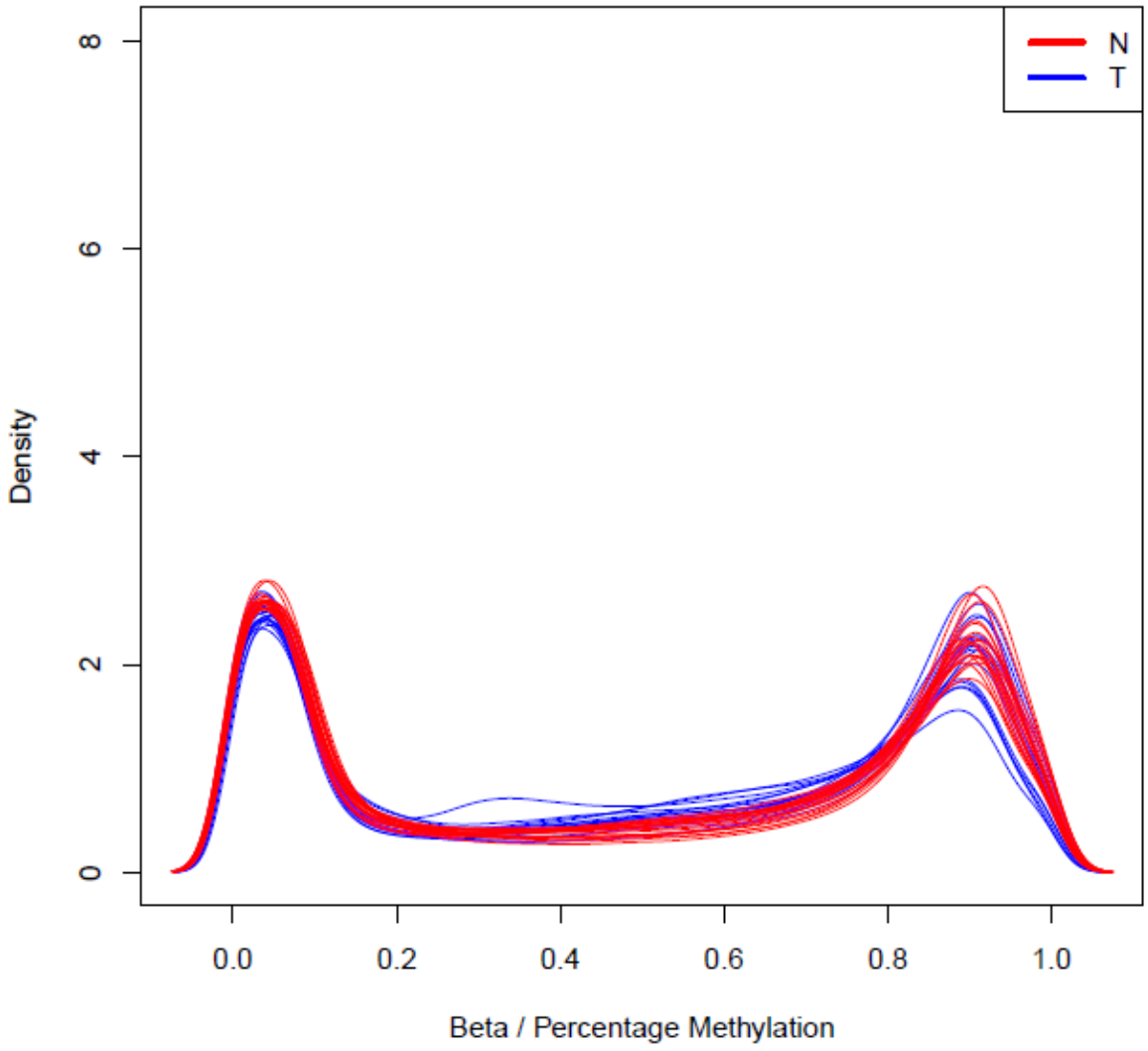

Figure 6: Histogram of beta values (Normal v/s Tumor)

5.3 Box plots for filtered Islands

The CpG islands were filtered with p-value < 0.05, fdr < 0.05 and beta value (methylation) > 0.2 and differentially methylated regions (DMRs) were identified and mapped to the corresponding gene location. A total of 453 DMRs were identified and box plots were created to visualize the difference in methylation. Some of the box plots are shown below.

5.4 Methylation v/s Expression plots

The filtered islands mapped to the corresponding genes were then integrated with the RNA-Seq data normalized expression values and matched with their TCGA IDs. Pearson correlation analysis was then carried out. 180 significant correlations were identified. Some of the methylation v/s expression plots are shown below.

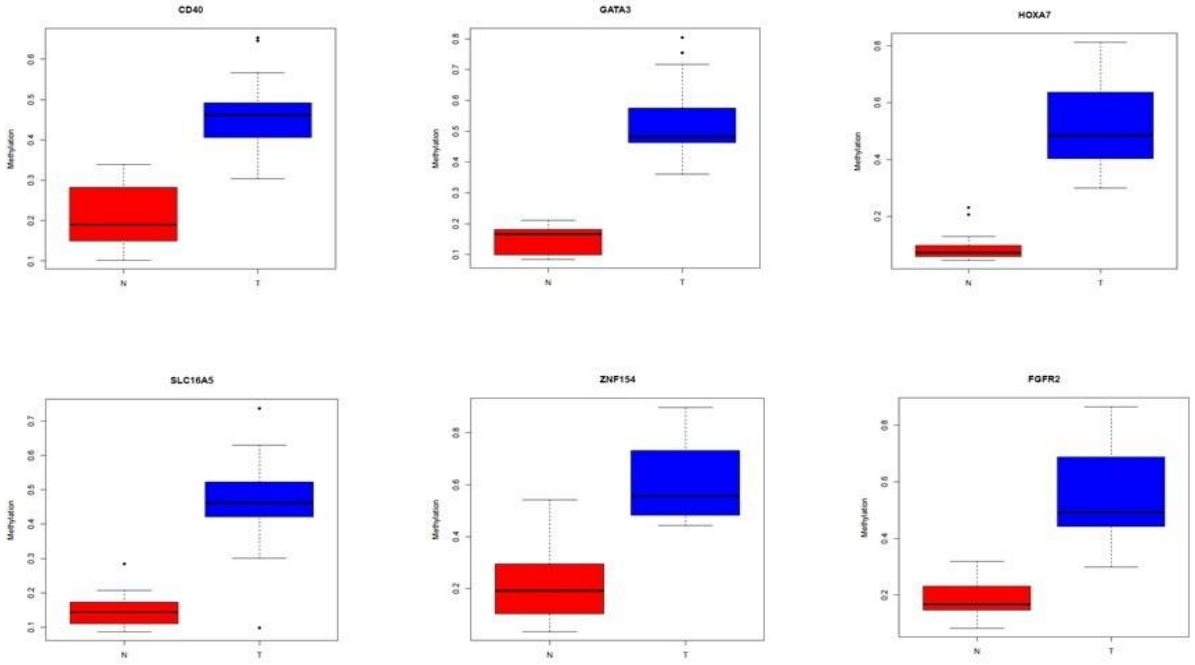

Figure 7: Box plots of methylation in various genes

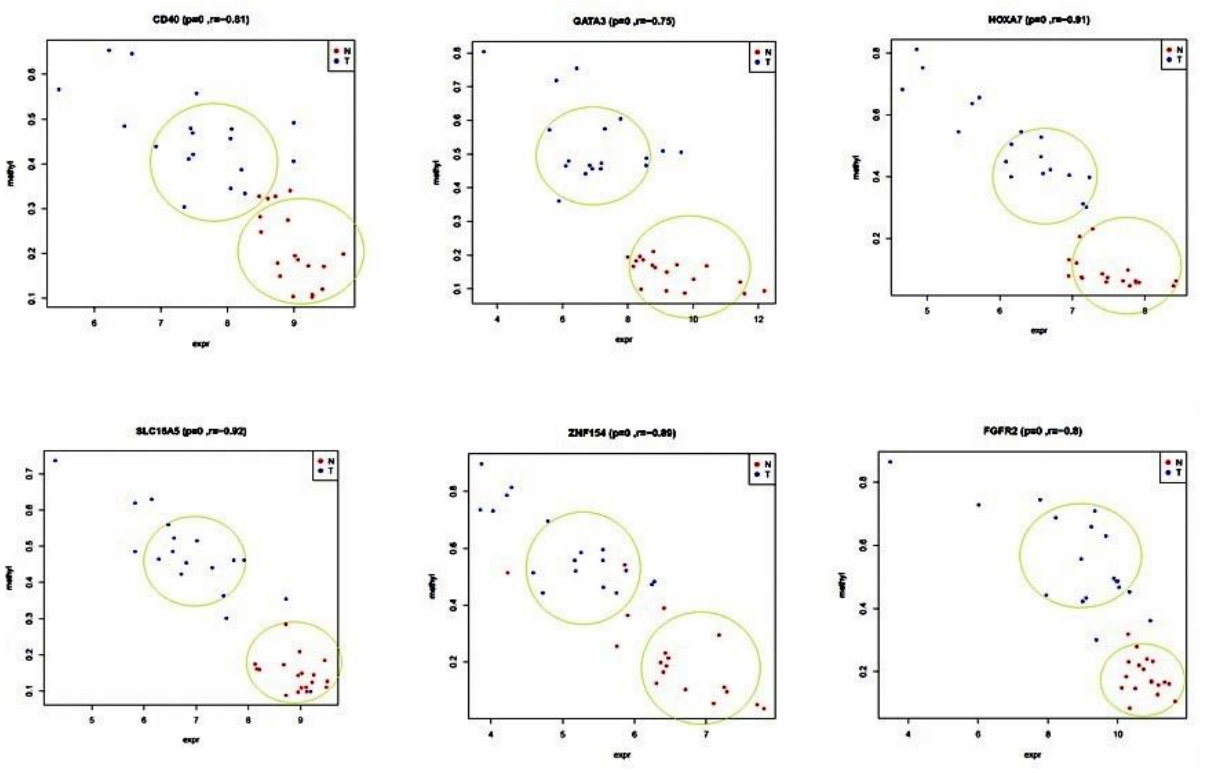

Figure 8: Correlation plots

5.5 Visualization in Integrative Genomics Viewer (IGV)

The COHCAP pipeline generates .wig files for the methylation value in normal and tumor samples and for the difference observed in the methylation in normal and tumor samples. These files can be visualized with IGV, Broad Institute with reference human genome hg19. Further visual inspection was carried out to identify the regions where the filtered CpG site lied in the promoter region of the gene. This may help prove the hypothesis that methylation may indeed be involved in the regulation of the gene expression.

A snapshot of the gene HOXA7 and CD40 is shown below. Here multiple CpG sites with a significant change in methylation can be seen all lying before exon 1of the gene HOXA7 and CD40.

Figure 9: Viewing CpG islands in HOXA7 gene

Figure 10: Viewing CpG islands in CD40 gene

5.6 Functional enrichment and pathway analysis

The genes filtered after integrating the methylation and normalized gene expression data were then fed to KEGG database and DAVID database and functional relevance was identified in terms of Gene ontology terms, involved disease or metabolic pathways or assigned biological function. The genes obtained after running correlation analysis between normalized methylation values and normalized expression values were filtered by visual inspection.

A total of 74 genes were uploaded to the DAVID database and functional information was obtained. Gene names were extracted for all the official symbols. 27 clusters were obtained with maximum enrichment score of ~1.86 with the GO (Gene Ontology) terms referring to extracellular region, metabolic processes, electron carrier activity, response to hormones like estrogen and steroid, immune system and transport.

A chart of KEGG pathways was also extracted for the genes with the default settings of DAVID.

Category	Term	Count	%	PValue	Genes
KEGG_PATHWAY	hsa00480:Glutathione metabolism	4	3.603604	0.006845	GSTM2, GPX3, GPX7, GSTP1
KEGG_PATHWAY	hsa00982:Drug metabolism	4	3.603604	0.012377	GSTM2, AOX1, MAOB, GSTP1
KEGG_PATHWAY	hsa05416:Viral myocarditis	4	3.603604	0.017818	CAV1, HLA-A, CD40, HLA-DOA
KEGG_PATHWAY	hsa05330:Allograft rejection	3	2.702703	0.032121	HLA-A, CD40, HLA-DOA
KEGG_PATHWAY	hsa00380:Tryptophan metabolism	3	2.702703	0.038996	AOX1, MAOB, HAAO
KEGG_PATHWAY	hsa05320:Autoimmune thyroid disease	3	2.702703	0.060406	HLA-A, CD40, HLA-DOA

KEGG_PATHWAY	hsa04510:Focal adhesion	5	4.504505	0.071558	CAV2, CAV1, CCND2, FLNA, PRKCB
KEGG_PATHWAY	hsa04514:Cell adhesion molecules (CAMs)	4	3.603604	0.084309	HLA-A, NRXN1, CD40, HLA-DOA

Table 6: KEGG pathway enrichment

A table containing functional information regarding gene name, Gene Ontology terms, proteins domains, KEGG pathways and OMIM diseases of all the existing genes in the database was downloaded from DAVID. The information was obtained for a total of 109 genes. The downloaded results are provided in the supplementary data.

Island	Gene	Cor < -0.5	p.value < 0.001	fdr < 0.001
chr17:73083866-73084495	SLC16A5	-0.922609302	1.25823E-15	3.88794E-13
chr7:27195601-27196567	HOXA7	-0.907957891	2.13721E-14	3.30199E-12
chr2:219646432-219647181	CYP27A1	-0.895283224	1.73318E-13	1.78517E-11
chr19:58220189-58220517	ZNF154	-0.891829163	2.92778E-13	2.26171E-11
chr19:39465846-39466403	FBXO17	-0.887159364	5.78705E-13	3.5764E-11
chr13:36919737-36921004	SPG20	-0.879996071	1.55621E-12	8.01447E-11
chrX:154842112-154842719	TMLHE	-0.878636462	1.86435E-12	8.22979E-11
chr1:53067880-53068608	GPX7	-0.865402638	9.74059E-12	3.67797E-10
chr11:67350928-67351953	GSTP1	-0.858137109	2.24554E-11	6.93872E-10
chr17:27044168-27045049	RAB34	-0.85645881	2.70564E-11	7.60038E-10
chrX:129243674-129245575	ELF4	-0.847509423	7.03641E-11	1.81188E-09
chr2:191044979-191045829	C2orf88	-0.838952105	1.66058E-10	3.66514E-09
chr12:81471569-81472119	ACSS3	-0.831700796	3.30929E-10	6.81713E-09
chr1:156674858-156676654	CRABP2	-0.828220114	4.55521E-10	8.79724E-09

chr6:2841810-2842273	SERPINB1	-0.826723596	5.21469E-10	9.47847E-09
chr2:56150340-56151180	EFEMP1	-0.817456946	1.17175E-09	2.01151E-08
chr20:44746822-44747060	CD40	-0.812882754	1.71871E-09	2.79516E-08
chr8:94712366-94713345	FAM92A1	-0.806716606	2.83442E-09	4.37917E-08
chr10:123356616-123358285	FGFR2	-0.80072774	4.53122E-09	6.66736E-08
chr2:21022564-21022934	C2orf43	-0.797045262	5.99971E-09	8.06098E-08
chr1:110210581-110210956	GSTM2	-0.796737064	6.14075E-09	8.06098E-08
chr7:116139774-116140352	CAV2	-0.796102299	6.44096E-09	8.06098E-08
chr20:3218578-3220930	SLC4A11	-0.795936005	6.52183E-09	8.06098E-08
chr5:150400000-150400490	GPX3	-0.793524556	7.80538E-09	9.23119E-08
chr14:23834435-23835947	EFS	-0.792856407	8.20033E-09	9.23119E-08
chr14:21492735-21494270	NDRG2	-0.792586872	8.36483E-09	9.23119E-08
chr3:38035701-38036000	VILL	-0.791180194	9.2741E-09	9.88172E-08
chrX:101906001-101907017	GPRASP1	-0.782488842	1.72514E-08	1.7769E-07
chr12:14720248-14721093	PLBD1	-0.778650403	2.24914E-08	2.20153E-07
chr2:26915603-26916551	KCNK3	-0.778451812	2.2799E-08	2.20153E-07
chr3:139257712-139257949	RBP1	-0.774543282	2.97021E-08	2.78119E-07
chr20:44098280-44099536	WFDC2	-0.774056228	3.06861E-08	2.78883E-07
chr16:31213566-31214287	PYCARD	-0.772423734	3.42086E-08	3.02013E-07
chr2:201450526-201451027	AOX1	-0.769315313	4.19698E-08	3.60241E-07
chr5:139283350-139284282	NRG2	-0.767029075	4.86835E-08	3.99122E-07
chr7:95025559-95026122	PON3	-0.766902414	4.9083E-08	3.99122E-07
chr7:30028518-30029822	SCRN1	-0.765534822	5.35934E-08	4.24625E-07
chr15:74658038-74658574	CYP11A1	-0.765129594	5.50017E-08	4.24888E-07
chrX:114468095-114468453	LRCH2	-0.75843796	8.37955E-08	6.31532E-07

chr10:8091374-8098329	GATA3	-0.753852189	1.10965E-07	8.16384E-07
chr2:43019665-43020509	HAAO	-0.748949606	1.48834E-07	1.04522E-06
chr1:48937304-48937683	SPATA6	-0.747401676	1.63066E-07	1.11972E-06
chr6:29716468-29717158	LOC285830	-0.746519942	1.71724E-07	1.15354E-06
chrX:56258951-56259159	KLF8	-0.745702745	1.80124E-07	1.18422E-06
chr14:75894308-75895469	JDP2	-0.742490162	2.16952E-07	1.39663E-06
chr10:99473084-99473291	MARVELD1	-0.741767938	2.26133E-07	1.42602E-06
chrX:100807793-100808048	ARMCX1	-0.737958026	2.80776E-07	1.73519E-06
chr14:24641053-24642220	REC8	-0.736735166	3.0074E-07	1.82213E-06
chr11:74178175-74178801	KCNE3	-0.733864469	3.52846E-07	2.05716E-06
chr7:116164703-116166735	CAV1	-0.733486237	3.60299E-07	2.06171E-06
chr16:88716989-88717606	CYBA	-0.73168265	3.9787E-07	2.21845E-06
chr14:23305893-23307013	MMP14	-0.731491857	4.02048E-07	2.21845E-06
chr6:146136325-146136564	FBXO30	-0.723116717	6.30633E-07	3.41869E-06
chr2:96990857-96991283	ITPRIPL1	-0.72074684	7.14213E-07	3.80503E-06
chr5:180017099-180019062	SCGB3A1	-0.718750375	7.92394E-07	4.04512E-06
chr20:55840216-55841794	BMP7	-0.718600923	7.98552E-07	4.04512E-06
chr6:117085739-117086942	FAM162B	-0.71711462	8.62226E-07	4.29722E-06
chr9:98783216-98784364	NCRNA00092	-0.715681078	9.28023E-07	4.55173E-06
chr1:150121695-150123078	PLEKHO1	-0.714479159	9.86709E-07	4.76395E-06
chr17:38599270-38599524	IGFBP4	-0.710851759	1.1851E-06	5.63376E-06
chr1:169396621-169396869	C1orf114	-0.709389413	1.27494E-06	5.96906E-06
chr10:79396095-79398495	KCNMA1	-0.706789395	1.45028E-06	6.68859E-06
chr2:71205563-71206529	ANKRD53	-0.705832653	1.52017E-06	6.90784E-06
chr2:29337983-29338909	CLIP4	-0.698858954	2.13065E-06	9.54162E-06

chr1:155947678-155948490	ARHGEF2	-0.697982705	2.22149E-06	9.80631E-06
chr11:61594996-61596710	FADS2	-0.697260126	2.29906E-06	1.00058E-05
chr11:12029737-12030841	DKK3	-0.692974162	2.81241E-06	1.19046E-05
chr5:156886969-156887440	NIPAL4	-0.686489702	3.79067E-06	1.58286E-05
chr13:43566074-43566508	EPSTI1	-0.685344149	3.9928E-06	1.64503E-05
chrX:73642494-73642766	SLC16A2	-0.682649143	4.50776E-06	1.83276E-05
chr4:41258759-41259867	UCHL1	-0.680532155	4.95406E-06	1.98806E-05
chr12:89745168-89748144	DUSP6	-0.676381224	5.94827E-06	2.35323E-05
chr3:46506161-46506416	LTF	-0.676120789	6.01634E-06	2.35323E-05
chr17:33700487-33700760	SLFN11	-0.674449286	6.47041E-06	2.4992E-05
chr14:51560116-51562487	TRIM9	-0.673913326	6.62249E-06	2.52636E-05
chr17:53342198-53343061	HLF	-0.673220329	6.82397E-06	2.57147E-05
chr5:115151348-115152713	CDO1	-0.671951956	7.20728E-06	2.65381E-05
chrX:134655102-134655750	DDX26B	-0.671929469	7.21425E-06	2.65381E-05
chr4:5709985-5710495	EVC2	-0.667417769	8.74366E-06	3.14364E-05
chr10:17270430-17272617	VIM	-0.66633547	9.15197E-06	3.25053E-05
chr3:149374709-149376300	WWTR1	-0.660916214	1.14703E-05	4.02762E-05
chr20:25565437-25566547	NINL	-0.660625129	1.16087E-05	4.03044E-05
chr17:7342829-7344028	FGF11	-0.656790905	1.35804E-05	4.6626E-05
chr6:29910202-29911367	HLA-A	-0.655790431	1.41427E-05	4.80229E-05
chr19:15344091-15344419	EPHX3	-0.652163122	1.63635E-05	5.49599E-05
chr11:111410932-111412199	LAYN	-0.645578509	2.12206E-05	7.0507E-05
chr1:95391837-95393116	CNN3	-0.641325545	2.50184E-05	8.21724E-05
chr4:16084195-16085735	PROM1	-0.641071834	2.52634E-05	8.21724E-05
chr15:89920793-89922768	LOC254559	-0.638941284	2.74078E-05	8.76192E-05

chr3:112051893-112052406	CD200	-0.632082657	3.54801E-05	0.000111871
chr22:38073037-38073412	LGALS1	-0.630268166	3.79486E-05	0.000118446
chr18:5543437-5544241	EPB41L3	-0.629810418	3.85954E-05	0.000118487
chrX:134124938-134125307	LOC644538	-0.629716845	3.87288E-05	0.000118487
chr17:33700989-33701657	SLFN11	-0.623587762	4.84356E-05	0.000146731
chr21:42798146-42798884	MX1	-0.621613879	5.20007E-05	0.000156002
chr8:22960384-22960927	TNFRSF10C	-0.619516933	5.60465E-05	0.000164937
chr1:203598471-203598853	ATP2B4	-0.619243051	5.65954E-05	0.000164981
chrX:153598874-153600604	FLNA	-0.618385261	5.83461E-05	0.000168495
chr3:44626325-44626794	ZNF660	-0.614219667	6.7564E-05	0.000193308
chrX:43741299-43741827	MAOB	-0.608189033	8.32449E-05	0.000235988
chr17:38333605-38334795	RAPGEFL1	-0.6049222	9.30452E-05	0.000259806
chr19:46800053-46800603	HIF3A	-0.604832538	9.33282E-05	0.000259806
chr14:69256676-69257036	ZFP36L1	-0.59412835	0.00013326	0.000364402
chr19:16186789-16188275	TPM4	-0.590400069	0.000150419	0.000407715
chr12:104850253-104852395	CHST11	-0.589456156	0.000155067	0.000416657
chr7:134143115-134144063	AKR1B1	-0.584867432	0.00017955	0.000478284
chr6:116691827-116692868	DSE	-0.581577136	0.000199184	0.000525375
chr12:4383193-4384405	CCND2	-0.581346604	0.000200629	0.000525375
chr2:50574045-50574817	NRXN1	-0.576420546	0.000233843	0.000607205
chr6:32975684-32975926	HLA-DOA	-0.57236633	0.000264784	0.000670641
chr2:235404502-235406541	ARL4C	-0.57115495	0.000274714	0.000690135
chr16:23846941-23848102	PRKCB	-0.566418794	0.000316812	0.000789475

Table 7: Filtered genes with corr < -0.5

CONCLUSION

Upon integration of expression data with methylation data, 180 significant correlations were obtained. On applying strict rules like correlation < -0.5, p value < 0.001 and false discovery rate < 0.001, 112 genes were filtered. The correlation plots were visually inspected and out of 112 genes, 74 genes were selected. These genes contain GSTP1 and FGFR2 which are known to play a role in Prostate cancer pathway (*KEGG Disease database*, Yang *et al*., 2013).

In the above table, all the 74 genes are highlighted and their corresponding CpG islands are provided in the supplementary files. Along with analysis files, functional enrichment files are also provided in the supplementary files.

In the IGV visual, it is clearly seen that the CpG sites lying upstream of the gene HOXA7 and CD40 are methylated. This does point to the fact that DNA methylation might indeed be responsible for the down-regulation of these genes. Further experimental validation must be carried out to actually correlate the roles of these significantly correlated genes to Prostate Adenocarcinoma. 74 genes which were stringently selected contained genes like GSTPI and FGFR2 (already known to play a role in prostate cancer pathway). The remaining genes can be projected as possible contenders playing a minor or a major role in prostate cancer progression and pathway. However choosing biomarkers out of these genes requires serious benchmarking and wet lab validation.

DISCUSSION

DNA methylation is known to play a critical role in regulating cellular functions and proceses (Robertson, 2005). Many human diseases have been linked to DNA methylation like cancers, neurodegenerative disorders and genetic disorders (Robertson, 2005; Suzuki and Bird, 2008).

DNA methylation and gene expression analysis carried out in Prostate Adenocarcinoma revealed genes whose expression was significantly correlated with DNA methylation in the CpG islands lying in the respective gene location. Some of these genes like GSTP1 and FGFR2 are already known to play a role in prostate cancer pathway.

Quality control was carried out for the 18 tumor matched with normal samples and 18 normal matched with tumor samples of Prostate Adenocarcinoma for DNA methylation data and clustering was represented in principal component analysis scatter plots. A dendogram was generated which clearly clusters normal samples together and tumor samples together. This clustering can also be clearly visualized in the principal component analysis scatter plot (Figure 4 and 5). This can be a possible application to predict the type of sample given its methylation data by analysing which category of cluster it falls in. Likewise subtyping of the tumors can be carried out which may prove to be of great help in diagnosis and healthcare.

This analysis is limited by the fact that along with off-shelve data from TCGA (The Cancer Genome Atlas), wet lab validation experiments are very important to support the interpretations made from analysing the DNA methylation data and gene expression data. It would be too early to comment on the topic of DNA methylation information used as biomarkers to subtype the various types of cancers as the research is still on-going and developing. However, it can be said that great responsibility lies in the hands of epi-geneticists and molecular biologists as these are the times for epigenetics.

FUTURE PERSPECTIVE

With the advent of 450K human methylation arrays, researchers are investing time and money in carrying out methylation analysis. There is a sudden explosion of data in the public data centres (Rakyan *et al.*, 2011). Innovative methodologies must be employed to make sense of the data.

This study marks as a proof of concept for integrating the studies of different platforms. TCGA has enabled the researchers to carry out a multi-platform analysis as it provides multi-platform data for same samples. The data at TCGA is well maintained and the architecture is well documented (https://wiki.nci.nih.gov/display/TCGA/).

In the current research, DNA Methylation and RNA-Seq data is integrated to analyse the correlation of methylation and expression. In future, other platforms can also be integrated to this research like somatic variations, clinical data and this type of study can be extended to other diseases and clustering can be carried out at a systems level to understand the relationship between genes, proteins, somatic mutations, and epigenetic modifications at a systems level.

This model of study will certainly add to the progress in the field of diagnosis and healthcare. Upcoming technologies like genome editing and personalized medicine require accurate and precise information which can be obtained by carrying out integrative investigation studies and working on to reduce the unknown parameters that contribute in disease–disease, disease–gene, protein–protein, metabolic and disease pathways.

Furthermore a disease model can be generated with integrated information and can be validated by experiments. Despite numerous efforts and investments, there is still a wide gap in the understanding of epigenetic modifications. We are far from understanding the reasons behind the differences observed in DNA methylation. Currently research aims at epigenetic therapies, predictive power of epigenetics and understanding developmental biology but a lot has to be done before one conquers this unfathomable journey (Bock, 2012).

REFERENCES

Abate-Shen C, Shen MM. (2000) Molecular genetics of prostate cancer. Genes Dev 14:2410-34 (2000)

Aryee MJ, Jaffe AE, Corrada-Bravo H, Ladd-Acosta C, Feinberg AP, Hansen KD, and Irizarry RA. (2014) Minfi: A flexible and comprehensive Bioconductor package for the analysis of Infinium DNA Methylation microarrays. Bioinformatics

Baylin, Stephen B (2005). DNA methylation and gene silencing in cancer. Nature Clinical Practice Oncology (2005) 2, S4-S11

Bird, Adrian (2007). Perceptions of epigenetics. NATURE|VOL 447| 24 MAY 2007|doi:10.1038/nature05913

Bock, Christoph (2012). Analysing and interpreting DNA methylation data. NATURE REVIEWS | GENETICS | VOLUME 13 | OCTOBER 2012

Chu, Yongjun; Corey, David R. (2012). RNA Sequencing: Platform Selection, Experimental Design, and Data Interpretation. NUCLEIC ACID THERAPEUTICS. Volume 22, Number 4, 2012.

Dong, JT; Lamb, PW; Rinker-Schaeffer, CW; Vukanovic, J; Ichikawa, T; Isaacs, JT; Barrett, JC. (1995). KAI1, a metastasis suppressor gene for prostate cancer on human chromosome 11p11.2. Science 12 May 1995: Vol. 268 no. 5212 pp. 884-886. DOI: 10.1126/science.7754374

Egger, Gerda, Liang, Gangning; Aparicio, Ana; Jones, Peter A. (2004). Epigenetics in human disease and prospects for epigenetic therapy. NATURE|VOL 429| 27 MAY 2004.

Esteller Manel (2008). Epigenetics in Cancer. N Engl J Med 2008; 358:1148-59.

Feinberg, Andrew P.; Tycko, Benjamin (2004). The history of cancer epigenetics. Nature Reviews|Cancer Volume 4

Gentleman RC, Carey VJ, Bates DM, Bolstad B, Dettling M, Dudoit S, Ellis B, Gautier L, Ge Y, Gentry J, Hornik K, Hothorn T, Huber W, Iacus S, Irizarry R, Leisch F, Li C, Maechler M, Rossini AJ, Sawitzki G, Smith C, Smyth G, Tierney L, Yang JY, Zhang J. (2004) Bioconductor: open software development for computational biology and bioinformatics. Genome Biol. 5(10):R80.

Holliday, Robin (2006). Epigenetics A Historical Overview. Epigenetics 2006; Vol. 1 Issue 2, 76-80; April/May/June 2006.

Hsing AW, Chokkalingam AP. (2006). Prostate cancer epidemiology. Front Biosci. 2006 May 1;11:1388-413.

Huang DW, Sherman BT, Lempicki RA. Bioinformatics enrichment tools: paths toward the comprehensive functional analysis of large gene lists. Nucleic Acids Res. 2009;37(1):1-13.

Huang DW, Sherman BT, Lempicki RA. Systematic and integrative analysis of large gene lists using DAVID Bioinformatics Resources. Nature Protoc. 2009;4(1):44-57.

Jones, Peter A.; Laird, Peter W. (1999). Cancer epigenetics comes of age. nature genetics|volume 21|february 1999

Jones, Peter A.; Takai Daiya (2001). The Role of DNA Methylation in Mammalian Epigenetics. Science 293 , 1068 (2001). DOI: 10.1126/science.1063852.

Meaburn, E; Schulz, R(2012). Next generation sequencing in epigenetics: insights and challenges. Semin Cell Dev Biol. 2012 Apr;23(2):192-9. doi: 10.1016/j.semcdb.2011.10.010. Epub 2011 Oct 19.

Morgan, Hugh D.; Santos, Fatima; Green, Kelly; Dean, Wendy; Reik Wolf (2005). Epigenetic reprogramming in mammals. Human Molecular Genetics, 2005, Vol. 14, Review Issue 1

Nelson WG, De Marzo AM, Isaacs WB. (2003) Prostate cancer. N Engl J Med 349:366-81 (2003).

Porkka KP, Visakorpi T. (2004) Molecular mechanisms of prostate cancer. Eur Urol 45:683-91 (2004)

Rakyan VK, Down TA, Balding DJ, *et al.* (2011) Epigenome-wide association studies for common human diseases. Nat Rev Genet; 12:529–41.

Riddihough, Guy and Zahn, Laura M. (2010). What is Epigenetics? Science 29 October 2010: Vol. 330 no. 6004 p. 611 DOI: 10.1126/science.330.6004.611

Robertson, K. (2005), DNA methylation and human disease. Nature Reviews Genetics 6, 597–610 (2005) doi: 10.1038/nrg1655

Robertson, K. D. (2002), DNA methylation and chromatin: Unraveling the tangled web. Oncogene 21, 5361–5379 (2002) doi:10.1038/sj.onc.1205609

Robinson JT, Thorvaldsdóttir H, Winckler W, Guttman M, Lander ES, Getz G, Mesirov JP. (2011) Integrative genomics viewer. Nat Biotechnol. 29(1):24-6

Robinson, James T.; Helga Thorvaldsdóttir; Wendy Winckler; Mitchell Guttman; Eric S. Lander; Gad Getz; Jill P. Mesirov. Integrative Genomics Viewer. Nature Biotechnology 29, 24–26 (2011)

Simmons, D. (2008), Epigenetic influence and disease. Nature Education 1(1):6

Suzuki, Miho M. and Bird, Adrian (2008). DNA methylation landscapes: provocative insights from epigenomics. NATURE REVIEWS | GENETICS JUNE 2008 | VOLUME 9

Thorvaldsdóttir, Helga; Robinson, James T.; Mesirov, Jill P.; Integrative Genomics Viewer (IGV): high-performance genomics data visualization and exploration. Briefings in Bioinformatics 14, 178-192 (2013).

Varambally, Sooryanarayana; Saravana M. Dhanasekaran; Ming Zhou; Terrence R. Barrette; Chandan Kumar-Sinha; Martin G. Sanda; Debashis Ghosh; Kenneth J. Pienta; Richard G. A. B. Sewalt; Arie P. Otte; Mark A. Rubin & Arul M. Chinnaiyan. (2002).The polycomb group protein EZH2 is involved in progression of prostate cancer. Nature 419, 624-629 (10 October 2002) | doi:10.1038/nature01075

Wang, Zhong; Gerstein, Mark; Snyder, Michael (2009) RNA-Seq: a revolutionary tool for transcriptomics. NATURE REVIEWS|GENETICS|VOL 10| JANUARY 2009

Warden CD, Lee H, Tompkins JD, Li X, Wang C, Riggs AD, Yu H, Jove R, Yuan YC. (2013) COHCAP: an integrative genomic pipeline for single-nucleotide resolution DNA methylation analysis. Nucleic Acids Res. 41 (11): e117

Yang F, Zhang Y, Ressler SJ, Ittmann MM, Ayala GE, Dang TD, Wang F, Rowley DR. (2013). FGFR1 is essential for prostate cancer progression and metastasis. Cancer Res. 2013 Jun 15; 73(12):3716-24. doi: 10.1158/0008-5472.CAN-12-3274. Epub 2013 Apr 10.

APPENDIX 1

#Pre-processing .idat files (DNA methylation 450k) using minfi

```
==== start code====
library(minfi)
idat.folder <- "path_to_idat_folder"
RG.raw <- read.450k.exp(idat.folder)
methyl.norm <- preprocessIllumina(RG.raw, bg.correct = TRUE, normalize = "controls")
beta.table <- getBeta(methyl.norm)
probes <- rownames(beta.table)
output.table <- data.frame(SiteID=probes, beta.table)
beta.file <- "minfi.txt"
write.table(output.table, file=beta.file, sep="\t", quote=F, row.names=F)
====end code====
```

#COHCAP analysis

```
==== start code====
library(COHCAP)
sample.file <- "COHCAP_sample_description.txt"
project.folder <- getwd()
project.name <- "COHCAP_folder_name"
beta.table <- COHCAP.annotate(beta.file, project.name, project.folder, platform="450k-UCSC")
COHCAP.qc(sample.file, beta.table, project.name, project.folder)
====end code====

filtered.sites <- COHCAP.site(sample.file, beta.table, project.name, project.folder, ref="N",
methyl.cutoff = 0.3)

filtered.islands <- COHCAP.avg.by.island(sample.file, filtered.sites, beta.table, project.name,
project.folder, ref="N", methyl.cutoff = 0.3)
```

#Pre-processing expression data

```
exp.normGenesFolder <- "E:/PROJECT/PROSTATE/DATA/3ee4b1b5-648d-4d2b-ab30-
b4c611cdaebe/RNASeqV2/UNC__IlluminaHiSeq_RNASeqV2/Level_3/"
exp.normFiles <- list.files(exp.normGenesFolder,pattern = "genes.normalized_results$",recursive =
TRUE,ignore.case = TRUE,full.names = TRUE )
for (file in exp.normFiles){
        if(!exists("exp.norm.counts")){
                exp.norm.counts <- read.table(file , header = TRUE)
        }
        else{
                temp_dataset <- read.table(file, header = TRUE)
                exp.norm.counts <- cbind(exp.norm.counts, temp_dataset["normalized_count"])
                rm(temp_dataset)
        }
 }
exp.norm.counts.M <- as.matrix(exp.norm.counts)
exp.logplusone.counts <- log2(exp.norm.counts.M  + 1)
write.table(exp.logplusone.counts, file="expression.txt", row.names = F, sep = "\t")
```

#COHCAP integration

```
COHCAP.integrate.avg.by.island(filtered.islands, project.name,project.folder, expression.file,
sample.file)
```

SUPPLEMENTARY MATERIAL

1. PRAD_DAVID_KEGG_RESULTS.xlsx
    a. Sheet 1: Name of the genes
    b. Sheet 2: Clusters of genes based on gene ontology terms and protein domains
    c. Sheet 3: Table containing all the information from DAVID
    d. Sheet 4: Table containing KEGG pathways
2. Box plots of 453 genes filtered based on differential methylation.
3. Scatter plots of 180 significantly correlated genes based on Methylation vs Expression data.
4. Table of 309 genes after integration of expression and methylation data along with correlation coefficient, p value and false discovery rate value.
5. Wiggle files for methylation in normal, tumor and difference in methylation to be viewed in a genome browser.
6. Table of significantly methylated CpG sites.
7. Table of significantly methylated CpG islands.