Methylated or miRNA-ed: Deciphering gene regulation networks in cancer

Maulik R. Kamdar[1]
[1]**Stanford Center for Biomedical Informatics Research, Stanford University**

Abstract

Cancer, a complex genetic disease, is classified into various categories, each of which is characterized by abnormal cell growth in a specific tissue. DNA methylation has been widely studied lately for its role in the regulation of expression of numerous cancer-specific genes. Dysregulation in the expression of miRNAs, a group of small non-coding RNA molecules that act as post-transcriptional regulators of gene expression, is also observed in several cancer typologies. Furthermore, DNA Methylation itself may cause miRNA dysregulation. In this paper, I develop a computational approach to analyze DNA Methylation, mRNASeq and miRNASeq data extracted from tumor samples of eight distinct cancer typologies in The Cancer Genome Atlas in conjunction, to understand gene regulatory networks in cancer. Preliminary results indicate that the proportion of differentially expressed protein-coding genes, under the effect of DNA methylation or miRNA dysregulation varies widely across different tumor typologies. However, I also find substantial overlap between the differentially expressed genes and their regulation indicating shared underlying mechanisms. This set of combined features consisting of mRNA sequences, DNA methylated regions and miRNA sequences could help us understand the molecular mechanisms underlying cancer, determine additional subcategories and lead to prediction of new diagnostic markers and therapeutic targets.

Introduction

Cancer, as we now know, is a complex genetic disease caused due to the involvement of multiple factors and numerous dysfunctional regulation networks. DNA hyper-methylation and hypo-methylation in promoter regions of the human genome have been known to play a key mediatory role in regulating gene expression of various tumor-suppressor and oncogenes (silencing or opening up) respectively, thereby affecting several cellular processes like cell signaling, apoptosis and immune system processes, and leading to a diseased state[7, 10]. On the other hand, expression of certain microRNAs (miRNAs), small non-coding RNA molecules (~22 nucleotides), is known to cause mRNA knockout or post-transcriptional regulation of gene expression. Several studies have shown that dysregulated miRNA expression could alleviate the initiation and progression of different cancer typologies[6, 8]. Similar to DNA methylation, there are two sides of miRNA expression, where down-regulation of certain miRNAs, which act as tumor suppressors and knockout oncogenes, has also been observed. For example, hsa-mir-375 is differentially expressed during breast lobular neoplasia[9], whereas down-regulation of hsa-mir-let7 is thought to activate the Ras oncogenes signaling pathway[15]. Now, there is new experimental evidence that DNA methylation is crucially involved in the dysregulation of miRNAs in cancer[21, 23]. Hence to get a better understanding of the molecular mechanisms underlying cancer as a whole, it is imperative to study gene regulation in context of DNA methylation and dysregulated miRNA expression. This could guide us towards advanced diagnostics and precise therapeutic treatments[2].

Related Work

There have been various attempts in the recent years, both experimental and computational to analyze DNA methylation and miRNA expression in conjunction with the mRNA expression to understand the underlying molecular mechanisms in cancer. Wang et al. extensively reviewed the aberrant methylation of miRNAs in the pathogenesis of lymphoid malignancies including chronic lymphocytic leukemia, multiple myeloma and acute lymphoblastic leukemia[23]. Suzuki et al. undertook a literature review to characterize the different kinds of miRNA interactions with genes, and screening methods to detect methylation of these miRNAs[21]. They also proposed the replacement of tumor-suppressive miRNAs as a potential therapeutic method. Huang et al. developed MethHC, a database comprising of a large collection of DNA methylation and mRNA/microRNA expression profiles in human cancer[13]. Research is also being undertaken to support the hypothesis that the underlying mechanisms between distinct cancer typologies may typically be similar. Hoadley et al. performed multiplatform analysis on the

molecular classification of 12 distinct cancer types, and found that subsets of Lung squamous, Head and Neck and Bladder cancer types could coalesce into one single type characterized by its specific alterations[11].

Identifying the gene regulatory networks underlying different cancer typologies, through an integrated, holistic analysis of different types of molecular datasets, could have various potential benefits. These combined features could be used to determine newer subcategories of cancer typologies help determine potential diagnostic biomarkers (tumor suppressor miRNAs or methylated regions) and help develop precise therapeutic approaches[2]. Most of the research to unravel the molecular mechanisms underlying cancer are focused on one particular typology, or are limited in their approach. In this project, I aim to develop an automated computational approach to identify dysfunctional gene regulation networks in different cancer typologies using the TCGA molecular datasets.

The specific aims of this research are three fold:

- An automated computational approach to determine the set of differentially expressed genes, cancer-specific methylated regions and miRNA sequences from The Cancer Genome Atlas molecular datasets, and to understand the correlation between the activities of these regions across different cancer typologies.
- To check if there is any underlying similarity between the different cancer typologies subjectively by intuitively determining shared features and objectively by analyzing the extent of overlap in the differentially expressed genes in their tumor samples through a Heatmap visualization.
- To extract the direct children ('Super Nodes') of the Gene Ontology *Biological Process* Node by reasoning, and performing Gene Ontology Enrichment analysis on the differentially expressed and methylated genes using these Super Nodes, to determine shared processes and typology-specific processes.

Materials and Methods

Datasets

The Cancer Genome Atlas[24] is a pilot project started in 2005 by the National Cancer Institute (NCI) and the National Human Genome Research Institute (NHGRI). The goal of this project is to catalog the genomic alternations found in all cancers. The TCGA public data portal gives open access to the cancer patient data and enables researchers to perform and validate their analysis on real data. The TCGA data portal has made available data pertaining to the molecular information (mRNA Expression, DNA Methylation, Single Nucleotide Polymorphisms (SNP), miRSeq, microarrays etc.) and clinical attributes (histology, follow up, medications, survival etc.) of around 9000 patients across 30 different cancer typologies in MySQL and Tab-separated formats. This data is available in a raw format with actual counts as observed in the sequencing arrays, or in a preprocessed version with normalized RPKM (Reads Per Kilobase per Million mapped reads) and Beta values for the expression and DNA Methylation respectively[19]. For this work, I have downloaded the mRNASeq, Methylation, miRNASeq and Clinical datasets for 8 cancer typologies – *Breast Invasive Carcinoma (BRCA), Bladder Urothelial Carcinoma (BLCA), Colon Adenocarcinoma (COAD), Colorectal Adenocarcinoma (COADREAD), Head and Neck Squamous Cell Carcinoma (HNSC), Lung Adenocarcinoma (LUAD), Lung Squamous Cell Carcinoma (LUSC)* and *Prostrate Adenocarcinoma (PRAD)*. I decided to work with this select set of typologies, based on their relevance in cancer research and the availability of adequate data to carry out my analysis.

CellBase provides a comprehensive collection of relevant biological information, like genes, proteins, regulatory elements and functional annotation, from various heterogeneous data sources[4]. For this work, I only download the list of protein-coding genes from *CellBase*. I also download pertinent information related to miRNA sequences in the human genome from two data sources. *miRBase* is a searchable database of published miRNA sequences and annotation[17]. Each entry in the *miRBase* Sequence database represents a predicted hairpin portion of a miRNA transcript, with information on the location and sequence of the mature miRNA sequence. The metadata of the miRNAs (chromosome locations, identifiers and mature/transcript information) for the human genome was downloaded. *miRTarBase* maintains a curated, exhaustive list of experimentally detected miRNA interactions with several gene targets in different species[12]. Even though both these data sources are also exposed under *CellBase* as REST APIs, I decided to download static versions for offline use. Finally, I downloaded the *Gene Ontology (GO)* and the GO functional term annotations for each gene from the *Gene Ontology Consortium*[1].

Preprocessing of the TCGA datasets

The *Broad Institute GDAC Firehose*[25] provides a methylation preprocessor which filters the raw TCGA methylation data for both Illumina HumanMethylation27k and Illumina HumanMethylation450k chips for use in downstream pipelines. The first stage of this pipeline involves excluding those rows where the gene symbol is NA, or more than 5% of the samples have NA beta values. Moreover, redundant columns containing probe metadata (gene symbol, chromosome and genomic coordinate) is similar across samples, so all columns except the first three are excluded. The probes that measure two gene markers at a time are split into individual rows, and then sorted according to *Hugo Gene Nomenclature (HGNC)* gene symbols. This stage was also similar for the preprocessing of *Illumina HiSeq RNASeq* Gene Expression and *Illumina HiSeq miRNASeq* datasets.

The second stage of this pipeline includes summarization of the methylation values of multiple probes on a gene basis level. For a given gene, all the methylation probes with a standard deviation below a specified threshold of 0.2 of the maximum standard deviation of a probe are discarded. Methylation values of multiple probes for one gene are correlated with the corresponding gene expression value, and only the methylation that is most anti-correlated with a gene is included. These usually exist within the promoter regions of the gene (*-1000 to +0.1*gene_length region*). The preprocessing pipeline also provides other modes of summarization like average, maximum and clinically correlated, but for this work, I used the gene expression correlated methylation values.

I further limit this study to include only the protein-coding genes, and the $\log_2$ normalize the RPKM expression values of the mRNASeq and miRNASeq datasets[20]. However, as I discuss further, my algorithm can intake any type of preprocessed files, including raw counts file and generate the list of significant differentially expressed and methylated genomic regions, as well as the significant paired correlations between the features.

Extracting Differentially Expressed and Methylated Cancer-specific Regions:

TCGA datasets have patient identifiers for both the tumor and the normal tissue samples, but the total number of tumor samples is more. Hence we identify only the set of those patients that have both normal and tumor samples for a given dataset uploaded in TCGA, by looking at the TCGA patient identifiers. The sample type (*tumor* or *normal*) is contained within the TCGA identifier value[26] of the sample, e.g. TCGA-02-0001-**01**C-01D-0182-01. Tumor types range from 01-09 and normal types range from 10-19. My algorithm intakes each of the three primary TCGA datasets (Methylation, Gene Expression and miRNASeq), along with the TSV formatted files of other auxiliary datasets, and generates a list of most significant, differentially expressed or methylated genes by a *paired T-Test* over the tumor and normal tissue samples of a patient for each gene. A paired T-Test was a logical choice, as we have at least 50-paired samples for all the eight cancer typologies that we are testing here.

To limit the false positives in the set of the differentially expressed protein-coding genes, I perform a *Bonferroni* correction on the computed P-values and threshold by 0.005 for each of these hypotheses. I further check the distribution of values in all the tumor samples, against those of normal samples, and remove those genes that have at least 10% of their tumor sample expression values within the range of $\text{Mean}_{\text{normal}} \pm \text{StdDev}_{\text{normal}}$. The user can adjust these settings as parameters to my script, and he can also customize whether to apply these corrections across all the three datasets or ignore them altogether.

For the set of significant differentially expressed protein-coding genes and the differentially methylated cancer-specific regions, I ascertain which of these cDMRs actually are the promoter regions of the protein-coding genes. I isolate the set of tumor samples, for which all the three molecular characteristics were measured and I run a *Pearson Correlation Test* between the beta values of the methylation of the tumor samples and the corresponding gene expression values. I repeat this process to find correlation between all the possible pairs of differentially expressed miRNAs and the genes and compute the *Pearson Correlation* statistic. I also check *miRTarbase* if any of the significant differentially expressed mRNAs and miRNAs are already known. Finally, I check which of the miRNAs are under the influence of DNA hyper- or hypo-methylation. Based on the shared set of differentially expressed genes, I develop Heatmap visualization to show the possible underlying degree of similarity between different cancer typologies.

Gene Ontology Super Nodes enrichment:

Once the set of significant differentially expressed and methylated protein-coding genes was obtained, I perform GO overrepresentation analysis, and abstract the retrieved GO terms to indicate the higher level biological processes, in which the significant genes are implicated. I extract the list of *GO Biological Process (GO BP)* functional annotations associated with each protein-coding gene. I reason over the *Gene Ontology* to find the set of immediate

children (Super Nodes) under the *GO Biological Process Term GO:0008150*. I associate all the *GO BP* functional annotations for a gene, to a *GO BP* Super Node and calculate a weighed score based on the T-test statistic to check which Super Nodes are generally enriched, and are there any cancer typologies that look outliers intuitively from a Heatmap visualization.

Results

I obtained a list of 20078 Human protein-coding genes from *CellBase*, along with 4694 miRNAs (hairpin and mature) from *miRBase*, between which I retrieved a total of 36278 known interactions from *miRTarbase*. Adding the miRNA transcripts to the list of genes, the total number of genomic regions analyzed in this work was 21959. After retrieving the required preprocessed files from the Firehose website on the 8 tumor typologies, I run the mRNASeq, miRSeq and the Methylation datasets through my algorithm, to retrieve a list of significant differentially expressed and cancer-specific differentially methylated genomic regions. The Statistics are shown in the Table 1. As multiple miRNAs may execute simultaneous post-transcriptional regulation of a gene, and one miRNA could regulate multiple genes, we see the large number of significant correlations between mRNA and miRNA differentially expressed regions, especially in Bladder Urothelial Carcinoma (BLCA) and Prostrate Adenocarcinoma (PRAD).

Table 1: Statistics indicating the number (#) of differentially expressed and methylated regions in tumor samples across multiple cancer typologies and the correlation between the different molecular characteristics.

Tumor	# Differentially Expressed protein-coding mRNAs	# Differentially Methylated promoter regions	# Differentially Expressed miRNAs	# Correlated Methylation and mRNA Expression	# Correlated mRNA and miRNA expression	# Correlated Methylation and miRNA expression
BRCA	1309	8575	183	819	2193	17
BLCA	192	4486	144	74	18424	10
COAD	779	6139	0	5	0	0
COADREAD	720	6858	0	4	0	0
HNSC	337	6779	186	137	893	12
LUAD	680	5689	123	340	657	9
LUSC	1945	7848	188	937	2301	13
PRAD	110	7082	154	72	11304	10

After determining those genes whose promoter regions were significantly hyper-methylated or hypo-methylated, I executed Pearson's correlation on the tumor samples that were common across all the three molecular types, to find if there was any statistically significant correlation between the two. In Table 2, are listed the top 10 genes for each cancer typology, which are either down-regulated or up-regulated (indicated in red and green respectively) and hyper-methylated or hypo-methylated (indicated using ^ and * symbols besides the gene identifier) in tumor samples, as compared to normal. Most of these genes are expected and I found high overlap with the *PAM50 BRCA gene list* and the *Cancer Gene Census*. For instance, there is a lot of ongoing research in studying the expression of proteins in the family of *matrix metalloproteinases* (MMPs), and therapeutic applications as potential biomarkers in breast cancer[5, 18]. As expected, hyper-methylation leads to silencing of tumor-suppressor genes whereas hypo-methylation leads to increased expression of oncogenes. As seen from Table 2, there are instances where a gene has been hyper-methylated but has increased expression, for example COL11A1 in COAD and KIF23 in LUSC. These cases are a part of a coordinated overexpression of certain genes in specific variants under the hypothesis *'shared "core" metastasis-associated gene expression'*, which are also being used to design biomarkers based on multi-invasion mechanisms[16]. Figure 1 shows a comparative box plot of the gene expression and methylation of gene DLG2 and PYCR1 in HNSC and LUAD tumor and normal samples respectively. It can also be seen that some genes like PYCR1 are shared between different cancer typologies (here LUAD and PRAD), indicating shared molecular basis of some subtypes of these typologies.

Table 2: Top 10 genes up-regulated or down-regulated (green and red respectively) as a direct consequence of

hypo-methylation or hyper-methylation (* and ^ respectively)

BRCA	BLCA	COAD	COADREAD	HNSC	LUAD	LUSC	PRAD
COL10A1*	PER1^	KRT80*	BEST4^	DLG2^	PYCR1*	FAM83B*	AMACR*
MMP11*	NR4A1^	ESM1*	KIAA1257*	GPT2^	RTKN2^	ESAM^	PYCR1*
ADAMTS5^	F10^	FOXQ1*	FOXQ1*	ARTN*	KANK3^	KIF23^	HIST3H2A*
KLHL29^	APOLD1^	BEST4^	KIAA1199*	GPD1L^	RAMP2^	TMEM88^	ETNK2^
SPRY2^	GSTM5^	ETV4*	ETV4*	NFIX^	LIMS2^	S1PR1^	DOK4^
HOXA4^	KLF2^	KIAA1199*	ESM1*	FAM3D^	WWC2^	SLC2A1*	CTF1^
NEK2*	HAAO^	TEAD4*	GLTP^	EXT1*	ACTN2^	PLK1*	APOBEC3C^
PER1^	FXYD1^	WNT2*	CLDN1*	SERPINH1*	TAL1^	PRC1*	MARCKSL1*
TPX2*	CSGALNACT1^	GLTP^	TEAD4*	FAM107A^	CLEC3B^	C15orf42*	CRYAB^
PPP1R12B^	ERAL1*	CLDN1*	WNT2*	QARS^	EPAS1^	CLEC3B^	EPHA10*
FAM126A^	PHYHIP^	COL11A1^	MTHFD1L*	INHBA*	FABP4^	CDCA5*	AOX1^

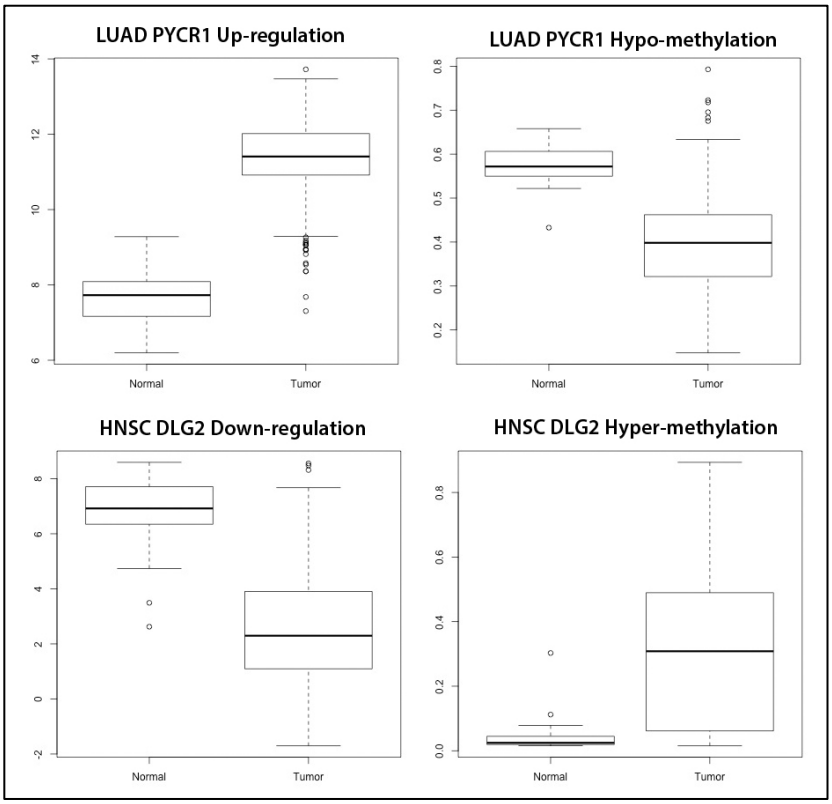

Figure 1: The up-regulation of PYCR1 gene in LUAD tumor samples due to hypo-methylation and the down regulation of the DLG2 gene in HNSC tumor samples due to the hyper-methylation in the promoter regions

Next I determined what is the proportion of the differentially expressed genes that are significantly methylated or are regulated by miRNAs. I had determined the above correlations using Pearson's correlation statistic, and the percentages are shown in Figure 2. It is noteworthy to see that 100% of the genes in some cancer typologies could be potentially regulated by miRNAs (BLCA and PRAD), whereas there was no such regulation determined in COAD and COADREAD cancers. This can also be due to the fact of organ-specific nature of these tissue samples. It was really strange to find that the proportion of differentially expressed genes that maybe methylated was also low in these tumor typologies, considering I found 6139 and 6858 potential cDMRs respectively. For a set of miRNAs

acting in any tumor typology, I generate a histogram to show how many protein-coding genes has it been correlated. Figure 3 indicates the high prevalence of miRNAs that could potentially regulate 30 or more genes simultaneously in the BLCA and the PRAD tumor typologies.

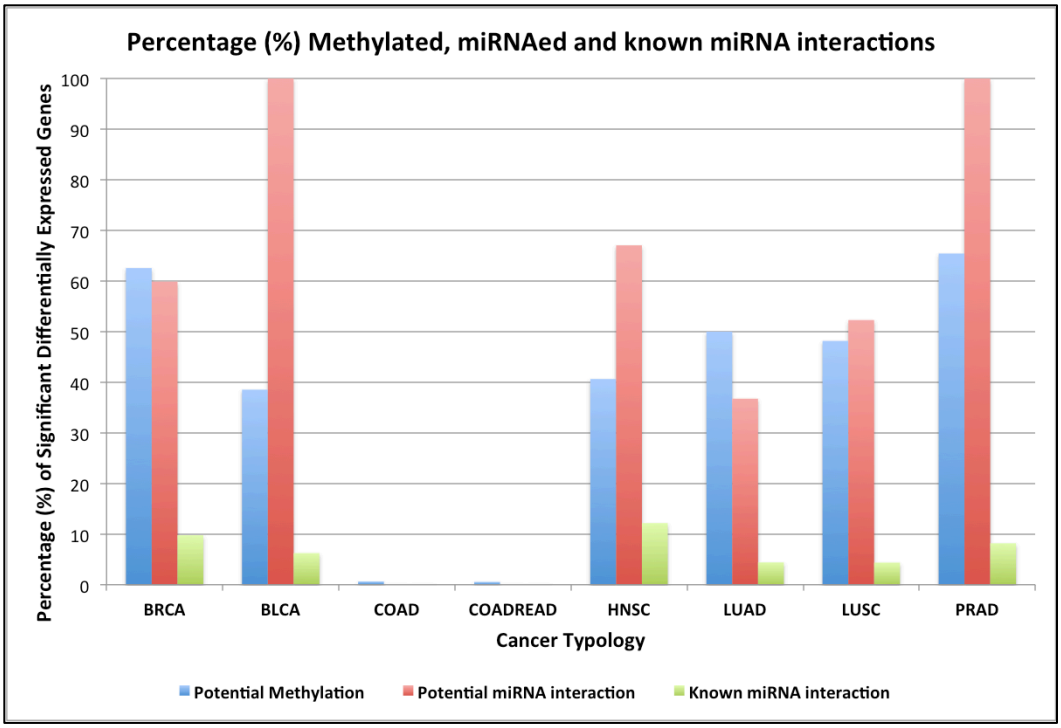

Figure 2: Percentage of significant differentially expressed genes that may be potentially methylated, regulated by miRNAs and are known miRNA interactions from *miRTarbase* across the different cancer typologies

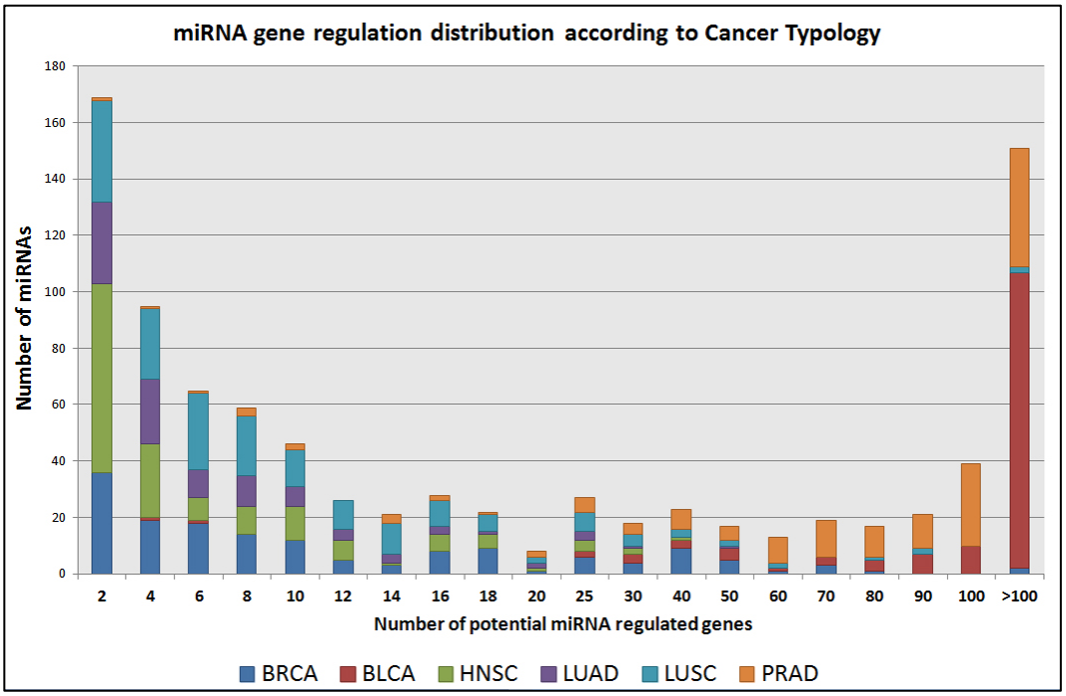

Figure 3: Number of miRNAs distributed according to the number of potential genes it may regulate

It has been shown that DNA Methylation may itself be involved in the dysregulation of miRNA expression, leading to downstream dysregulation and post-transcriptional modification of mRNA expression. I also determine the hyper-

methylation and hypo-methylation of these miRNA regions, as the miRNA transcripts are also measured under the *Illumina HumanMethylation450k* chips. I found significant association between these two molecular characteristics, and in Table 3, I display a selected number of miRNA sequences (mature and hairpin) that are differentially expressed under the influence of DNA methylation. I also show the total number of samples, where the methylation was observed and simultaneous up-regulation or down-regulation was also observed. Finally, it is worth noticing that these miRNAs may be influential in more than one cancer typology under the same regulation. Through a brief literature review, I was able to find significant evidence pertaining to the oncogenic role of the first two miRNAs *hsa-mir-425* and *hsa-mir-17* in different cancer typologies[22, 9]. Hence our method is able to obtain substantial results, which could be relevant for further analysis. Like with mRNA expression, I also found instances namely *hsa-mir-let7* and *hsa-mir-589* in BRCA that were hyper-methylated yet up-regulated.

Table 3: Regulation of miRNA expression under DNA methylation, as observed in various cancer typologies

miRNA Transcript	Methylation Status	% samples	miRNA	Expression Status	% samples	Cancer Typologies
MIR425	hypomethylated	92.23	hsa-mir-425	upregulated	89	BRCA, BLCA, LUAD, PRAD
MIR17	hypomethylated	58.84	hsa-mir-17	upregulated	88.66	BLCA, HNSC, LUAD, PRAD
MIR345	hypermethylated	59.66	hsa-mir-345	upregulated	53.15	BLCA, HNSC, LUAD, LUSC
MIR19B1	hypomethylated	83.01	hsa-mir-19b-1	upregulated	74.08	BLCA, LUAD, PRAD
MIR20A	hypomethylated	83.01	hsa-mir-20a	upregulated	87.04	BLCA, LUAD, PRAD
MIR19A	hypomethylated	83.01	hsa-mir-19a	upregulated	91.44	BLCA, HNSC, PRAD
MIR125B1	hypermethylated	81.63	hsa-mir-125b-1	downregulated	70.55	BRCA, HNSC, LUSC
MIR205	hypomethylated	77.18	hsa-mir-205	upregulated	65.04	BLCA, HNSC, LUSC
MIR15B	hypomethylated	62.03	hsa-mir-15b	upregulated	73.11	BRCA, BLCA
MIR16-2	hypomethylated	62.03	hsa-mir-16-2	upregulated	83.62	BRCA, BLCA
MIR18A	hypomethylated	58.84	hsa-mir-18a	upregulated	61.05	HNSC, PRAD
MIR301A	hypomethylated	64.05	hsa-mir-301a	upregulated	71.97	BRCA, LUSC
MIR324	hypermethylated	71.59	hsa-mir-324	upregulated	80.88	BRCA, HNSC

Based on the differentially expressed genes, I tried to determine pairs of cancer typologies that share the same set of genes, thus indicating a possible underlying mechanism similarity. This is visualized using Heatmap visualization as shown in Figure 4. The color of each cell is determined from the quantity – (Total number of common genes*100/Total number of differentially expressed genes in the 'row' cancer typology). The very limited set of differentially expressed genes in the Prostrate Adenocarcinoma (110) as compared to others is obvious from the plot. It is noteworthy that COAD and COADREAD basically have >75% of common differentially expressed genes, as well as LUAD and LUSC have >50%. An extended visualization involving the different subcategories could lead to meta-clustering of different cancer typologies based on co-occurring genes.

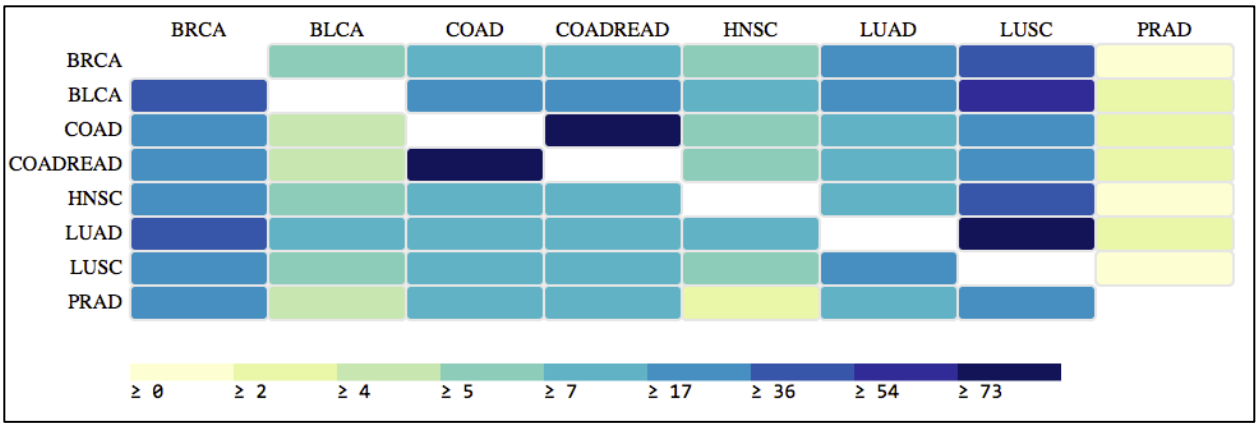

Figure 4: Similarity between different cancer typologies based on the common differentially expressed genes.

Finally, I conduct *Gene Ontology* Enrichment analysis, based on the differentially expressed and methylated protein-coding genes in each cancer typology, and abstract the results to display only the Super Nodes in the *Biological Process* ontology of GO. It can be seen in the Figure 5, that the GO Super Nodes *Biological Regulation (GO:0065007), Cellular Process (GO:0009987)* and *Single organism Process (GO:0044699)* are significantly enriched across all the cancer typologies, which was obvious as the underlying set of differentially expressed and methylated genes should be somehow involved in important cellular regulatory processes. Even as most of the enriched terms remain consistent across the typologies, there are some specific outliers. For example, it can be seen that differentially methylated genes in COAD and COADREAD are significantly involved in *Signaling (GO:0023052)* and *Biological Adhesion (GO:0022610)*, whereas differentially expressed genes in LUAD, BRCA and LUSC are significantly enriched with terms under *Immune System Process (GO:0002376)*. It has been shown previously that cell-cell adhesion pathways and integrin signaling is important for apoptosis of colon cancer cells[3, 14].

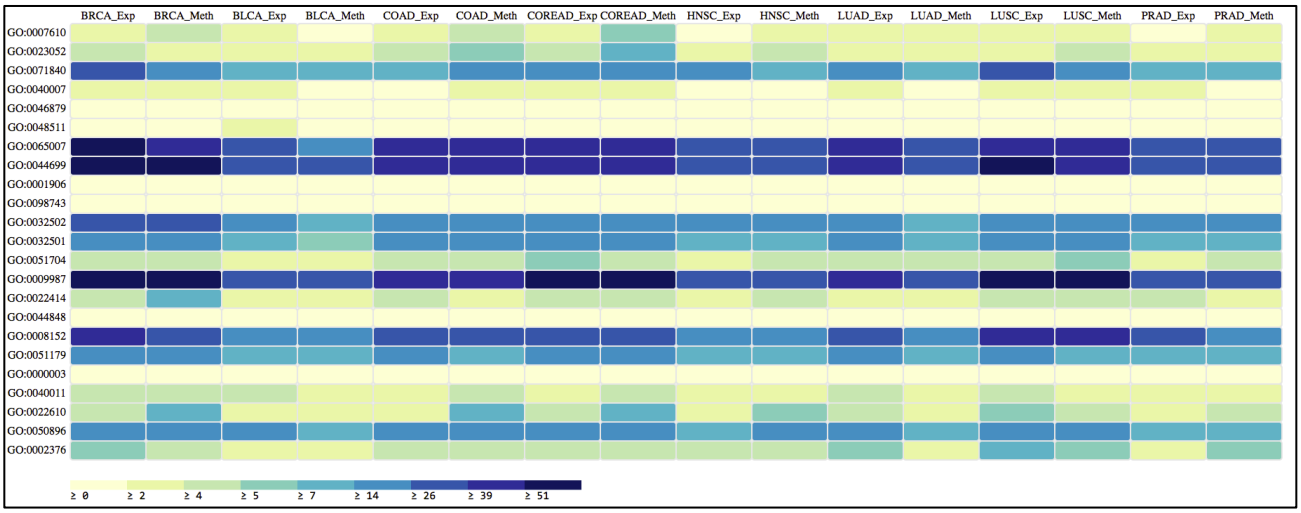

Figure 5: Heatmap visualization of the Gene Ontology Biological Process Super Nodes that are associated with the differentially expressed and methylated genes across different cancer typologies.

Discussion

Increasing numbers of DNA Methylation and mRNA/miRNA expression profiles are being published every day since the advent of high-throughput gene sequencing technologies. Hence automated tools that analyze this data using user-provided customizations are required. 'Data analysis' has already replaced 'Data generation' as the crucial step in precision medicine. It was noteworthy to see from the observations that the differentially expressed genes and correlated methylation and miRNA expression profiles were actually implicated in some underlying molecular mechanisms of cancer. I also observed that while the general hypothesis is that a hyper-methylated gene

is down-regulated, I actually saw instances where this is not true, both in the case of mRNA and miRNA expression. From the results, it is also evident that while the modes of regulation may vary between tumor typologies, different cancers may actually have subsets that have significantly overlapping mechanisms.

A more rigorous evaluation of the method described in this paper is required. Combination of molecular features (namely expression and methylation) derived from The Cancer Genome Atlas datasets can be used in supervised learning algorithms to predict the histology subcategory of the tumor. It is also possible to start from this set of features and feed them in unsupervised machine learning algorithms like K-Means or Spectral Biclustering, to predict new subcategories and clusters. Optimum number of k-clusters could be determined using the number of significant features obtained from an ANOVA test across samples in different clusters. A decreasing curve is generated as the cluster size (k) increases and the user can choose the optimum threshold. Moreover, it would be worthwhile to include other type of molecular datasets like SNP and microarrays that are also made available under TCGA to actually understand gene regulation in the full context. Overlaying this information on molecular pathways like KEGG or Reactome could lead to an increased understanding of Cancer. I would like to modify my P-value threshold selection statistic from *Bonferroni* correction to *False Discovery Rates*, as *Bonferroni* corrections are often considered conservative and we might lose some true positives. That being said, my experience with DESeq2 and other R packages resulted in a very large number of differentially expressed and methylated regions (~10,000). It would be easy to replace this statistical mechanism in my algorithm.

Conclusion

In this paper, I present a computational approach to analyze the mRNASeq, DNA Methylation and the miRNASeq datasets of The Cancer Genome Atlas in conjunction to understand how differentially expressed genes between tumor and normal samples are regulated. I extend my analysis across 8 different cancer typologies, to find that gene expression is regulated both through DNA Methylation and miRNA post-transcriptional regulation; hence dysfunction in either of these regulatory networks leads to up-regulation and down-regulation of certain genes. Also, my approach found that DNA Methylation could also cause dysfunctional miRNA expression. The mode of regulation varies across different cancer typologies, as I found that both Bladder Urothelial Carcinoma and Prostrate Adenocarcinoma show 100% potential interactions between miRNAs and mRNAs, whereas 0% is detected in Colon Adenocarcinoma and Colorectal Adenocarcinoma. My findings suggest that there is substantial overlap between the differentially expressed genes and its regulation in specific cancer typologies hinting at shared underlying mechanisms. I also perform Gene Ontology Enrichment Analysis on the differentially expressed genes, but limit myself to only present the direct children of the *Biological Process* Node. This set of combined features that includes mRNAs, miRNAs and DNA methylated regions could be used to perform unsupervised learning for subclassification, understanding the molecular mechanisms of cancer and determine new diagnostic markers and therapeutic targets.

References

1. Ashburner, Michael, Catherine A. Ball, Judith A. Blake, David Botstein, Heather Butler, J. Michael Cherry, et al. "Gene Ontology: tool for the unification of biology." *Nature genetics* 25, no. 1 (2000): 25-29.
2. Ashour, Nadia, Javier C. Angulo, Guillermo Andrés, Raúl Alelú, Ana González‐Corpas, María V. Toledo, José M. Rodríguez‐Barbero, Jose I. López, Manuel Sánchez‐Chapado, and Santiago Ropero. "A DNA hypermethylation profile reveals new potential biomarkers for prostate cancer diagnosis and prognosis." *The Prostate* 74, no. 12 (2014): 1171-1182.
3. Avizienyte, Egle, Anne W. Wyke, Robert J. Jones, Gordon W. McLean, M. Andrew Westhoff, Valerie G. Brunton, and Margaret C. Frame. "Src-induced de-regulation of E-cadherin in colon cancer cells requires integrin signalling." *Nature cell biology* 4, no. 8 (2002): 632-638.
4. Bleda, Marta, Joaquin Tarraga, Alejandro de Maria, Francisco Salavert, Luz Garcia-Alonso, Matilde Celma, Ainoha Martin, Joaquin Dopazo, and Ignacio Medina. "CellBase, a comprehensive collection of RESTful web services for retrieving relevant biological information from heterogeneous sources." *Nucleic acids research* (2012): gks575.

5. Cheng, Chun-Wen, Jyh-Cherng Yu, Hsiao-Wei Wang, Chiun-Sheng Huang, Jia-Ching Shieh, Yi-Ping Fu, Chia-Wei Chang, Pei-Ei Wu, and Chen-Yang Shen. "The clinical implications of MMP-11 and CK-20 expression in human breast cancer." *Clinica chimica acta* 411, no. 3 (2010): 234-241.
6. Croce, Carlo M. "Causes and consequences of microRNA dysregulation in cancer." *Nature reviews genetics* 10, no. 10 (2009): 704-714.
7. Das, Partha M., and Rakesh Singal. "DNA methylation and cancer." *Journal of Clinical Oncology* 22, no. 22 (2004): 4632-4642.
8. Esquela-Kerscher, Aurora, and Frank J. Slack. "Oncomirs—microRNAs with a role in cancer." *Nature Reviews Cancer* 6, no. 4 (2006): 259-269.
9. Giricz, Orsi, Paul A. Reynolds, Andrew Ramnauth, Christina Liu, Tao Wang, Lesley Stead, Geoffrey Childs et al. "Hsa‐miR‐375 is differentially expressed during breast lobular neoplasia and promotes loss of mammary acinar polarity." *The Journal of pathology* 226, no. 1 (2012): 108-119.
10. Hansen, Kasper Daniel, Winston Timp, Héctor Corrada Bravo, Sarven Sabunciyan, Benjamin Langmead, Oliver G. McDonald, Bo Wen et al. "Increased methylation variation in epigenetic domains across cancer types." *Nature genetics* 43, no. 8 (2011): 768-775.
11. Hoadley, Katherine A., Christina Yau, Denise M. Wolf, Andrew D. Cherniack, David Tamborero, Sam Ng, Max DM Leiserson et al. "Multiplatform analysis of 12 cancer types reveals molecular classification within and across tissues of origin." *Cell* 158, no. 4 (2014): 929-944.
12. Hsu, Sheng-Da, Feng-Mao Lin, Wei-Yun Wu, Chao Liang, Wei-Chih Huang, Wen-Ling Chan, Wen-Ting Tsai et al. "miRTarBase: a database curates experimentally validated microRNA–target interactions." *Nucleic acids research* (2010): gkq1107.
13. Huang, Wei-Yun, Sheng-Da Hsu, Hsi-Yuan Huang, Yi-Ming Sun, Chih-Hung Chou, Shun-Long Weng, and Hsien-Da Huang. "MethHC: a database of DNA methylation and gene expression in human cancer." *Nucleic acids research* (2014): gku1151.
14. Jaiswal, Aruna S., Benjamin P. Marlow, Nirupama Gupta, and Satya Narayan. "Beta-catenin-mediated transactivation and cell-cell adhesion pathways are important in curcumin (diferuylmethane)-induced growth arrest and apoptosis in colon cancer cells." *Oncogene* 21, no. 55 (2002): 8414-8427.
15. Johnson, Steven M., Helge Grosshans, Jaclyn Shingara, Mike Byrom, Rich Jarvis, Angie Cheng, Emmanuel Labourier, Kristy L. Reinert, David Brown, and Frank J. Slack. "RAS is regulated by the let-7 microRNA family." *Cell* 120, no. 5 (2005): 635-647.
16. Kim, Hoon, John Watkinson, Vinay Varadan, and Dimitris Anastassiou. "Multi-cancer computational analysis reveals invasion-associated variant of desmoplastic reaction involving INHBA, THBS2 and COL11A1." *BMC medical genomics* 3, no. 1 (2010): 51.
17. Kozomara, Ana, and Sam Griffiths-Jones. "miRBase: annotating high confidence microRNAs using deep sequencing data." *Nucleic acids research* (2013): gkt1181.
18. Merdad, Adnan, Sajjad Karim, Hans-Juergen Schulten, Ashraf Dallol, Abdelbaset Buhmeida, Fatima Al-Thubaity, Mamdooh A. Gari, Adeel GA Chaudhary, Adel M. Abuzenadah, and Mohammed H. Al-Qahtani. "Expression of matrix metalloproteinases (MMPs) in primary human breast cancer: MMP-9 as a potential biomarker for cancer invasion and metastasis." *Anticancer research* 34, no. 3 (2014): 1355-1366.
19. Mortazavi, Ali, Brian A. Williams, Kenneth McCue, Lorian Schaeffer, and Barbara Wold. "Mapping and quantifying mammalian transcriptomes by RNA-Seq." *Nature methods* 5, no. 7 (2008): 621-628.
20. Quackenbush, John. "Microarray data normalization and transformation." *Nature genetics* 32 (2002): 496-501.
21. Suzuki, Hiromu, Reo Maruyama, Eiichiro Yamamoto, and Masahiro Kai. "DNA methylation and microRNA dysregulation in cancer." *Molecular oncology* 6, no. 6 (2012): 567-578.
22. Takakura, Shu, Norisato Mitsutake, Masahiro Nakashima, Hiroyuki Namba, Vladimir A. Saenko, Tatiana I. Rogounovitch, Yuka Nakazawa, Tomayoshi Hayashi, Akira Ohtsuru, and Shunichi Yamashita. "Oncogenic role of miR‐17‐92 cluster in anaplastic thyroid cancer cells." *Cancer science* 99, no. 6 (2008): 1147-1154.
23. Wang, Lu Qian, Raymond Liang, and Chor Sang Chim. "Methylation of tumor suppressor microRNAs: lessons from lymphoid malignancies." (2012): 755-765.
24. The Cancer Genome Atlas Project (TCGA). https://tcga-data.nci.nih.gov/tcga/
25. Broad Institute GDAC Firehose. http://gdac.broadinstitute.org/
26. TCGA Barcode Identification. https://wiki.nci.nih.gov/display/TCGA/TCGA+barcode