Statistical and Bioinformatic Analysis of Hemimethylation Patterns in Non-Small Cell Lung Cancer

Shuying Sun (✉️ s_s355@txstate.edu)
Texas State University San Marcos  https://orcid.org/0000-0003-3974-6996

Austin Zane
Texas A&M University College Station

Carolyn Fulton
Schreiner University

Jasmine Philipoom
Case Western Reserve University

Research article

Keywords: Methylation, Hemimethylation, Lung Cancer, Bioinformatics, Epigenetics

DOI: https://doi.org/10.21203/rs.3.rs-17794/v2

License: Creative Commons Attribution 4.0 International License. Read Full License
Abstract

Background:

DNA methylation is an epigenetic event involving the addition of a methyl-group to a cytosine-guanine base pair (i.e., CpG site). It is associated with different cancers. Our research focuses on studying non-small cell lung cancer hemimethylation, which refers to methylation occurring on only one of the two DNA strands. Many studies often assume that methylation occurs on both DNA strands at a CpG site. However, recent publications show the existence of hemimethylation and its significant impact. Therefore, it is important to identify cancer hemimethylation patterns.

Methods:

In this paper, we use the Wilcoxon signed rank test to identify hemimethylated CpG sites based on publicly available non-small cell lung cancer methylation sequencing data. We then identify two types of hemimethylated CpG clusters, regular and polarity clusters, and genes with large numbers of hemimethylated sites. Highly hemimethylated genes are then studied for their biological interactions using available bioinformatics tools.

Results:

In this paper, we have conducted the first-ever investigation of hemimethylation in lung cancer. Our results show that hemimethylation does exist in lung cells either as singletons or clusters. Most clusters contain only two or three CpG sites. Polarity clusters are much shorter than regular clusters and appear less frequently. The majority of clusters found in tumor samples have no overlap with clusters found in normal samples, and vice versa. Several genes that are known to be associated with cancer are hemimethylated differently between the cancerous and normal samples. Furthermore, highly hemimethylated genes exhibit many different interactions with other genes that may be associated with cancer. Hemimethylation has diverse patterns and frequencies that are comparable between normal and tumorous cells. Therefore, hemimethylation may be related to both normal and tumor cell development.

Conclusions:
Our research has identified CpG clusters and genes that are hemimethylated in normal and lung tumor samples. Due to the potential impact of hemimethylation on gene expression and cell function, these clusters and genes may be important to advance our understanding of the development and progression of non-small cell lung cancer.

**Background**

Lung cancer is a leading cause of death in the United States; more patients die of lung cancer than of breast, prostate, and colon cancers combined. The American Cancer Society predicts that in 2020 alone there will be 228,820 new cases of lung cancer diagnosed and 135,720 deaths in the United States [1]. The five-year survival rate of lung cancer is much lower than many other prominent cancers such as breast, colorectal, and prostate, as only 19.4% of patients survive beyond five years of having the disease. The rate of survival can be as high as 57.4% when the cancer is still localized. However, the majority (57%) of patients are diagnosed in late stages, and the five-year survival rate of these patients is only 5.2% [2].

In order to diagnose and treat lung cancer, it is often important to identify genetic and epigenetic biomarkers. One important epigenetic biomarker or event is DNA methylation. It is the covalent bonding of a methyl group (-CH$_3$) to a CpG site in a mammalian cell; this is an epigenetic alteration to the DNA, meaning the DNA sequence does not change. A CpG site is the nucleotide base cytosine bonded to the base guanine by a phosphate (5’-CpG-3’) [3]. The correlation between methylation patterns on specific CpG sites and gene expression has been studied as methylation enhances or mutes particular genes [4]. DNA methylation patterns are maintained and changed mainly through two dynamic processes: methylation maintenance and de novo methylation [5, 6]. Methylation maintenance allows for preservation of methylation patterns across replication generations, maintaining valuable methylation levels. De novo methylation occurs on symmetrically unmethylated CpG sites and increases methylation levels over cell generations [5]. Demethylation is the action of a methyl group being removed from a CpG site, and it can be observed in two forms: passive and active [6]. Passive demethylation is an error during maintenance methylation, resulting in bare or hypomethylated CpG sites on the nascent strand, whilst the parent strand is methylated. In contrast, active demethylation is the deliberate removal of a methyl group from a CpG site [7].

Both demethylation and de novo methylation can lead to the development of hemimethylation, i.e., methylation occurring on only one DNA strand of a CpG site but not the other. Because demethylation and de novo methylation are related to the loss and gain of methylation respectively, hemimethylation, as a transitional state, may be associated with the changes of methylation patterns and levels as cancer progresses [7]. That is, hemimethylation may be closely related to abnormal hypermethylation and hypomethylation patterns found in a cancer genome. In fact, Ehrlich and Lacey find that the study of hemimethylation provides valuable insight into cancer-associated active demethylation and
hypomethylation [5]. Exactly how different hemimethylation patterns affect gene expression is not yet well documented, except for the recent findings by Xu and Corces, who show that the elimination of hemimethylation can reduce chromatic interactions [8]. They also show that in inner cell mass cells, there are a large number of hemimethylated CpG sites on gene bodies. These hemimethylated sites play a pleiotropic role on gene expression [8].

DNA methylation patterns and levels can vary as cancer progresses [7]. Abnormal hypermethylation and hypomethylation are commonly known characteristics of cancerous cells. Identification of these different states of methylation can assist in the detection of cancerous cells long before they would appear in clinical examinations or before symptoms are apparent. Hemimethylation as a transitional state or indicator of hypomethylation and hypermethylation can help medical researchers to monitor how far the disease has progressed. Knowledge of the hemimethylation can allow for better comprehension of certain cancers and provide better insight toward the development of treatment methods. Therefore, it is important to investigate it in cancer. Hemimethylation has been researched previously for breast cancer cell lines [9], but not yet for lung cancer. Lung cancer is a great candidate for this investigation as it is challenging to detect early-stage lung cancer. Its symptoms are often obscure or mistakable due to the consequence of previous smoking habits. Utilizing hemimethylated genes to identify carcinogenic development may be beneficial in lung cancer diagnosis. The purpose of this research study is to identify hemimethylation patterns in both normal and tumorous samples of non-small cell lung cancer patients using publicly available methylation sequencing data.

**Methods**

In this study, to identify hemimethylation patterns, we will analyze the methylation sequencing data generated from tumor and adjacent normal tissues of 18 male non-small cell lung cancer patients in their fifties to seventies. The reduced representation bisulfite sequencing datasets of these patients are publicly available [10]. Sequencing reads are aligned to the hg38 reference genome and methylation signals are obtained using the BRAT-bw software package [11]. All methylation datasets consist of the methylation signals of CpG sites. These methylation levels are calculated based on the original or raw number of reads, that is, the methylation level or ratio at each CpG site is calculated as the number of methylated reads divided by total number of reads. Further analysis is then performed on CpG sites with at least four methylation signals for both strands.

Hemimethylation is a particular kind of methylation pattern. If a CpG is methylated on the forward strand but not on the reverse strand, it is defined as a MU hemimethylation site. If a CpG is methylated on the reverse strand but not on the forward strand, it is defined as a UM hemimethylation site. If a CpG exhibits no significant hemimethylation, it is defined as an NS site. If no data is available to be analyzed at a CpG, that site is defined as an NA site. Hemimethylation occurs not only at solitary CpG sites, but also at consecutive ones, known as hemimethylation clusters. Such clusters can manifest in one of two distinct patterns: regular or polar [5, 7, 9, 12], see Figure 1. A regular cluster can be observed when sequential CpG sites are methylated on the same strand and unmethylated on the other. A polar or polarity cluster occurs...
when consecutive CpG sites are methylated on opposite strands. Next, we will explain in detail how to identify these different hemimethylation patterns.

The Wilcoxon signed rank test is utilized to investigate if hemimethylation exists at each CpG site [9, 13]. This particular test is selected instead of a regular statistical t-test because the independence and normality assumptions are not satisfied. The null hypothesis is that, at each CpG site, there is no methylation level difference between the forward (or positive) and reverse (or negative) strands. For every CpG site, there are two methylation signals, one from the forward and one from the reverse strand. That is, there are up to 18 pairs of methylation signals at each CpG site as there are 18 samples. For each pair, the forward strand methylation level is subtracted from the reverse strand methylation level. The absolute value of the difference and the sign of the difference (negative or positive) are recorded separately. Pairs with zero difference are discarded. The absolute differences of the pairs are then ranked from smallest to largest so that the pair with the smallest absolute difference is ranked one. Lastly, a test statistic is calculated by summing all of the ranks after multiplying them by their respective signs (i.e., signed-ranks). This test statistic follows a specific distribution, and it is evaluated using a reference table. If the test statistic falls into the rejection region that is determined by the critical value from the table, then the null hypothesis is rejected. The rejection decision means that there is difference between the forward and reverse methylation levels at a CpG site. If the null hypothesis is not rejected, it shows that there is not a significant difference. The Wilcoxon signed rank test is conducted for tumor and normal samples separately. That is, we identify hemimethylated CpG sites for tumor and normal samples separately and then compare them.

The test results are filtered based on two metrics: forward and reverse strand methylation mean difference and Wilcoxon test p-value. CpG sites with a large mean difference in methylation levels and a p-value that is less than 0.05 are identified as hemimethylated CpG sites. Significant CpG sites are annotated to show which gene promoter region or gene body they belong to. Additionally, clusters of CpG sites are identified as either regular or polar patterns. For example, MM-UU and MU-UM are regular and polar clusters respectively. MM-UU means at two consecutive CpG sites, methylation occurs on the positive strands (i.e., MM) but not on the reverse strand (i.e., UU). MU-UM means at two consecutive CpG sites, on the positive strand they are methylated (M) and unmethylated (U) respectively (i.e., MU), and on the reverse strand they are unmethylated (U) and methylated (M) respectively (i.e., UM). The CpG sites that are not in clusters are called singletons. The lengths of all clusters in tumor and normal cells are shown in histograms. The percentages of CpG sites in regular clusters, polar clusters, and singleton points are summarized. All these summarized results of tumor and normal samples are further compared using statistical tests. For those CpG sites in clusters, DNA strands in the tumor and adjacent normal cells are compared, and overlapping clusters are identified. We'll show the key results in the Results section.

**Results**

Hemimethylated CpG sites for both normal and tumor cells are identified using the Wilcoxon tests. Table 1 describes the proportions of hemimethylation sites that are in clusters depending on the p-value (p <
0.05) and three mean difference cutoff values. There are similar numbers of hemimethylation sites in tumor and normal samples, but the proportion in clusters is slightly higher in normal samples. When comparing the proportions between normal and tumor, we get the following three p-values, 0.00039, 0.00035, and 0.277, for the three mean difference cutoff values 0.4, 0.6, and 0.8 respectively. For the rest of this paper, our analysis will focus on the hemimethylation sites identified based on the p-value of 0.05 and the absolute mean difference greater than or equal to 0.4.

| Mean difference | Normal Total | Sites in clusters | Percentage | Tumor Total | Sites in clusters | Percentage |
|-----------------|--------------|------------------|------------|-------------|------------------|------------|
| ≥0.4            | 7351         | 1510             | 20.54%     | 7330        | 1336             | 18.23%     |
| ≥0.6            | 2588         | 348              | 13.45%     | 2743        | 282              | 10.28%     |
| ≥0.8            | 723          | 53               | 7.33%      | 823         | 49               | 5.95%      |

**Table 1:** Number of hemimethylated CpG sites and percentage of sites in clusters. Each row is for a mean difference level. The two panels (3 columns each) are for normal and tumor samples respectively.

Tumor and normal samples’ hemimethylation CpG sites are compared in Table 2. The first row of this table, i.e., the T.MU row, indicates the total number of MU hemimethylation CpG sites in tumor (T) cells. Among these sites, 1697 of them are also hemimethylated in normal cells (N.MU), 1688 of them are not significantly hemimethylated in normal (N.NS), and 217 of them have no data in normal cells (N.NA). The first column of Table 2, i.e., the N.MU column, shows the total number of MU hemimethylation CpG sites in normal (N) cells. Among these sites, 1697 of them are also hemimethylated in tumor cells (T.MU), 1728 of them are not significantly hemimethylated in tumor (T.NS), and 268 of them have no data in tumor cells (T.NA).

|       | N.MU | N.UM | N.NS | N.NA |
|-------|------|------|------|------|
| T.MU  | 1697 | 0    | 1688 | 217  |
| T.UM  | 0    | 1597 | 1892 | 239  |
| T.NS  | 1728 | 1789 | 1895429 | 101322 |
| T.NA  | 268  | 272  | 98209 | 27295013 |

**Table 2:** Comparison of normal and tumorous hemimethylation site patterns. Each row is for the tumor (T) sample and each column is for the normal (N) sample with various hemimethylation types. T.MU refers to CpG sites that are methylated (M) on the forward strand and unmethylated (U) on the reverse strand in tumor (T) samples. N.MU refers to CpG sites with the MU hemimethylation in normal (N) samples. T.NS and N.NS refer to CpG sites of a corresponding tissue type that are not significantly hemimethylated. Similarly, T.NA and N.NA refer to CpG sites that have no data for the given cell type.

Tumor and normal samples’ hemimethylation clusters are compared in Table 3. This table shows that most clusters only have two or three CpG sites and cluster frequency decreases with increased cluster length, meaning large congregations of hemimethylation are infrequent.
| Cluster Pattern | Normal | Tumor |
|-----------------|--------|-------|
| MMMMMMMMM-UUUUUU | 1      | 1     |
| MMMMMMMM-UUUUUUU | 1      | 1     |
| MMMMMU-UUUUU    | 2      | 2     |
| MMMMM-UUUUU     | 2      | 2     |
| MMMM-UUUUU      | 5      | 3     |
| MMMM-UUUU       | 6      | 7     |
| MMM-UUUU        | 18     | 13    |
| MMU-UUM         | 55     | 32    |
| MMU-UUU         | 168    | 153   |
| MU-UUM          | 0      | 1     |
| UMM-MUU         | 28     | 32    |
| UM-MUU          | 1      | 0     |
| UUM-MMU         | 7      | 4     |
| UU-MMU          | 1      | 0     |
| UU-MUM          | 195    | 172   |
| UUU-MMM         | 52     | 44    |
| UUUU-MMMM       | 22     | 22    |
| UUUUU-MM        | 9      | 14    |
| UUUUUU-MMMM     | 3      | 4     |
| UUUUUU-MM        | 0      | 1     |
| UUUUUUU-MMM     | 4      | 3     |
| UUUUUU-MMM-MMU  | 1      | 0     |
| UUUUUU-MM-MMM-MMU | 2   | 2     |
| Total           | 583    | 513   |

**Table 3**: Normal and tumor hemimethylation cluster patterns. The first column is the cluster pattern, separating forward and reverse strands by “-“. The second and third columns are the counts of such patterns in normal and tumor samples respectively.

The length of a cluster is defined as the total number of base pairs between the first and the last CpG sites in the cluster. Figure 2 shows 4 histograms of cluster lengths. These histograms display the length distributions of polarity patterns in tumor, polarity patterns in normal, regular patterns in tumor, and regular patterns in normal samples. Regular and polarity patterns are analyzed separately because polarity clusters tend to be much shorter. In fact, many of the polarity clusters are less than 40 base pairs long and a majority of them are less than 10 base pairs long (see peaks in the top panels of Figure 2). Many of the regular clusters are relatively short, i.e., less than 60 base pairs long, but a small amount of them are longer than that with a maximum length of around 100 to 120 base pairs. A Wilcoxon rank-sum test is performed to compare the difference between the lengths of clusters in normal and tumor cells. The test result is insignificant (p-value =0.12).
For the two main hemimethylation cluster patterns, regular cluster and polarity cluster, we summarize them in detail in Table 4 and Table 5. Table 4 describes the proportions of different regular clusters in normal and tumor DNA. Table 5 describes the proportions of different polarity patterns in normal and tumor DNA. Polarity clusters appear less frequently than regular patterns, as seen by the difference in the number of sites between Tables 4 and 5. For example, tumor samples have a total of 477 regular clusters and only 36 polar clusters.

One way to detect which clusters may be related to cancer is to compare the cluster locations between tumor DNA and normal DNA. Some clusters may appear in the same sites in both tumor and normal samples, but others may be found only in tumor or only in normal. In Figure 3, we show two typical hemimethylation clusters: one that is only identified in tumor DNA and one that is only identified in normal DNA. The first two pairs of bars represent two CpG sites in normal DNA. The second two represent two CpG sites in tumor DNA. We see in the first (or left) plot that there is a large difference between the forward and reverse strands in the tumor CpG sites, whereas the normal CpG sites are quite similar. This tells us that there is a cluster containing two CpG sites that is found only in tumor DNA. Similarly, the second (or right) plot describes a cluster that appears only in normal DNA. In fact, there is almost no methylation in the tumor reverse strands, while the normal reverse strands are almost fully methylated. The forward strand methylation levels are similarly low in tumor and normal DNA, so we observe normal-only hemimethylation in the two sites.

In order to study hemimethylation patterns thoroughly, we compare the 513 tumor clusters with the 583 normal clusters and summarize the results in Table 6. This table shows that multiple kinds of overlaps can be found between tumor and normal. Hemimethylation clusters that occur only in tumor or normal samples are shown in Column B. 695 (313 tumor only and 382 normal only) clusters fall into these categories, and these are the clusters or regions that may be associated with cancer. Column C counts the number of clusters that are exactly the same for normal and tumor. Column D indicates the situations in which a tumor cluster begins and ends within a normal cluster (i.e., tumor cluster contained within the bounds of a normal cluster), and vice versa as shown in Column E. For example, a tumor cluster’s start

| Regular Clusters | Normal | Tumor |
|------------------|--------|-------|
| MM-UU            | 168    | 153   |
|                  | 30.66% | 32.075% |
| UU-MM            | 195    | 172   |
|                  | 35.58% | 36.059% |
| Bigger cluster   | 185    | 152   |
|                  | 33.76% | 31.866% |
| Total            | 548    | 477   |
|                  | 100%   | 100%  |

Table 4: Regular clusters with corresponding percentages. Bigger clusters (see the fourth row) are the ones with 3 or more hemimethylated CpG sites.

| Polarity Clusters | Normal | Tumor |
|-------------------|--------|-------|
| MU-UM             | 28     | 32    |
|                   | 80%    | 88.89% |
| UM-MU             | 7      | 4     |
|                   | 20%    | 11.11% |
| Total             | 35     | 36    |
|                   | 100%   | 100%  |

Table 5: Polarity clusters with corresponding percentages.
and end positions on a chromosome are 150 and 170 base pairs. It is located within a normal cluster that has the start and end positions of 120 and 190 base pairs. Column D, which represents tumor clusters that are embedded in normal clusters, shows different counts for normal and tumor samples because there are two instances of multiple normal clusters located in one tumor cluster. Similarly, Column E, which represents normal clusters that are embedded in tumor clusters, shows different counts because there are three tumor clusters that are located in one normal cluster. Column F represents all other kinds of overlap. For example, there are two normal clusters that have some overlap with the same tumor cluster.

The second row of Table 6 shows that among the 513 tumor clusters, 313 of them belong to tumor only; 140 clusters also show up in normal samples; 25 tumor clusters are short ones and they are located within long normal clusters; 23 tumor clusters are long ones in which short normal clusters are located; and 12 tumor clusters are partially overlapped with normal clusters. The third row of Table 6 shows that among the 583 normal clusters, 382 of them belong to normal only; 140 clusters also show up in tumor samples; 23 normal clusters are long ones and they cover short tumor clusters; 25 normal clusters are short ones and they are located within long tumor clusters; and 13 normal clusters are partially overlapped with tumor clusters. A detailed version of Table 6 is shown in the Supplemental Table 1, in which the number of different clusters in each chromosome is listed for both tumor and normal samples.

| A   | B           | C         | D         | E         | F         |
|-----|-------------|-----------|-----------|-----------|-----------|
| Tumor Total 513 | Tumor Only 313 | Exact Overlap 140 | Tumor in Normal 25 | Normal in Tumor 23 | Other Overlap 12 |
| Normal Total 583 | Normal Only 382 | Exact Overlap 140 | Tumor in Normal 23 | Normal in Tumor 25 | Other Overlap 13 |

**Table 6:** Tumor and normal cluster comparison results. Columns are for different overlap (or non-overlap) patterns. The two rows are for tumor and normal, respectively.

After identifying hemimethylated CpG sites, we may also map them back to genes. That is, we provide the annotation for each CpG site by providing the gene name in whose gene body or promoter region a hemimethylation site is located. We call this analysis gene annotation and summarizing such will provide the frequency on how many hemimethylated CpG sites a gene has. This annotation analysis is important because highly hemimethylated genes may play an important role. Table 7 shows the frequency of hemimethylated CpG sites in gene bodies. Each column shows how many genes have \( n \) hemimethylated CpG sites in their gene bodies, where \( n \) is given in the first row. The second row describes the distribution for tumor genes and the third row describes the distribution for normal genes. Similarly, Table 8 describes the frequency of hemimethylated CpG sites in promoter regions. Table 7 displays that the large majority of gene bodies have at most 3 hemimethylated CpG sites in both tumor and normal samples, but a few have more than 10. When looking at promoter regions, Table 8 shows none have 10 or more and the large majority of genes have 1 or 2 hemimethylated CpG sites.

| No. of Hemimethylation sites per gene body | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | \( \geq 10 \) |
|-------------------------------------------|---|---|---|---|---|---|---|---|---|---------|
| Tumor                                    | 1133 | 250 | 79 | 37 | 17 | 4 | 2 | 0 | 4 |         |
| Normal                                   | 1118 | 229 | 73 | 32 | 114 | 31 | 1 | 1 | 5 |         |
Table 7: Hemimethylation frequency measured in gene bodies for both tumor and normal samples.

| No. of Hemimethylation sites per prom region | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 |
|-------------------------------------------|---|---|---|---|---|---|---|---|
| Tumor                                     | 223| 23 | 6 | 5 | 0 | 2 | 0 | 1 |
| Normal                                    | 256| 36 | 13| 3 | 2 | 1 | 1 | 0 |

Table 8: Hemimethylation frequency measured in promoter regions for both tumor and normal samples.

With the gene annotation analysis, we can identify genes that have relatively more hemimethylation sites. In particular, we select the genes that have at least 5 hemimethylation sites in tumor only, in normal only, and in both normal and tumor samples. These genes are summarized in Tables 9, 10, and 11 respectively. Note, there are not many genes with a large number of hemimethylated sites. Therefore, we choose a relatively small number (i.e., five) to find a reasonable number of genes that meet this criterion for us to do further analysis. In addition, the datasets used in this project are generalized using the reduced representation bisulfite sequencing method. For the method, only a small percent of the CpG sites in a genome are sequenced [12, 14]. If the methylation sequencing datasets used in this study are generated based on the whole genome bisulfite sequencing method, more hemimethylated CpG sites can be found in different genes.

There are 41 genes with the most hemimethylation in tumor DNA, see Table 9. Among these genes, TP73 [15-17], GNAS [18-22], and NOTCH1 [23, 24] are notable ones with known relations to cancer. Table 9 shows that among these 41 genes, one is a tumor suppressor (WT1), three are oncogenes (GNAS, NOTCH1, and PRDM16), and of those three, two are translocated cancer genes (NOTCH1 and PRDM16). There are also eight transcription factors in this table (HDAC4, IRX2, NFATC1, PRDM16, RUNX3, SIX3, TP73, and WT1). Table 10 shows 35 genes with the most hemimethylation in normal DNA. Among these genes, four are oncogenes (CBFA2T3, GNAS, PDGFB and PRDM16). Of the oncogenes, three are translocated cancer genes (CBFA2T3, PDGFB and PRDM16). There are also seven transcription factors in this table (CBFA2T3, HOXA3, IRX2, MEIS1, NFIC, PRDM16, and ZFPM1). Note that no tumor suppressor genes are hemimethylated in the normal cells. For genes belonging to two key gene families (i.e., transcription factor and oncogene), we have compared their proportions in tumor and normal samples using statistical tests. The test p-values are 0.96 for the transcript factor family and 0.54 for the oncogene family. There is no significant difference. Table 11 shows 36 genes with the most hemimethylation in both normal and tumor DNA. Among these genes, two are oncogenes and also translocated cancer genes (CBFA2T3 and PRDM16). There are also six transcription factors in this table (KLF5, HOXA2, CBFA2T3, HOXA3, ISL2, and PRDM16). All three gene tables have some transcription factor genes, which may affect the gene expression of other cancer-related genes that are not found to be hemimethylated.
| Gene name    | Count | Family | Gene Description                                           |
|-------------|-------|--------|----------------------------------------------------------|
| RGS14       | 17    | -      | regulator of G protein signaling 14                      |
| MEX3A       | 16    | -      | mex-3 RNA binding family member A                        |
| WT1         | 11    | TF, TS | WT1 transcription factor                                 |
| PRDM16      | 10    | OG, TF, TCG | PR/SET domain 16                                              |
| ZDHHC9      | 10    | -      | zinc finger DHHC-type containing 9                        |
| AGAP2       | 8     | -      | ArfGAP with GTPase domain, ankyrin repeat and PH domain 2 |
| GNAS        | 8     | OG     | GNAS complex locus                                       |
| EXOC3L2     | 8     | -      | exocyst complex component 3 like 2                       |
| PTPRN2      | 7     | -      | protein tyrosine phosphatase receptor type N2             |
| FANK1       | 7     | -      | fibronectin type III and ankyrin repeat domains 1         |
| UNC93B1     | 7     | -      | unc-93 homolog B1, TLR signaling regulator                |
| IGSF9B      | 7     | -      | immunoglobulin superfamily member 9B                     |
| GNAS-AS1    | 7     | -      | GNAS antisense RNA 1                                     |
| MAD1L1      | 7     | -      | mitotic arrest deficient 1 like 1                         |
| TSPAN9      | 7     | -      | tetraspanin 9                                            |
| PTPRM       | 7     | -      | protein tyrosine phosphatase receptor type M              |
| TP73        | 6     | TF     | tumor protein p73                                        |
| IFT140      | 6     | -      | intraflagellar transport 140                             |
| NFATC1      | 6     | TF     | nuclear factor of activated T cells 1                     |
| DGKA        | 6     | -      | diacylglycerol kinase alpha                               |
| FMNL1       | 6     | -      | formin like 1                                             |
| CACNA1I     | 6     | -      | calcium voltage-gated channel subunit alpha1 I           |
| LOC101927636| 6    | -      | RNA Gene affiliated with the lncRNA class                |
| HDAC4       | 5     | TF     | histone deacetylase 4                                    |
| IRX2        | 5     | TF, HP | iroquois homeobox 2                                      |
| ANKRD33B    | 5     | -      | ankyrin repeat domain 33B                                |
| LINC00537   | 5     | -      | Long Intergenic Non-Protein Coding RNA 537               |
| NOTCH1      | 5     | OG, TCG| notch receptor 1                                          |
| ANO2        | 5     | -      | anoctamin 2                                               |
| CACNA1H     | 5     | -      | calcium voltage-gated channel subunit alpha1 H           |
| RUNX3       | 5     | TF     | runt related transcription factor 3                       |
| SIX3        | 5     | TF, HP | SIX homeobox 3                                            |
| FZD7        | 5     | -      | frizzled class receptor 7                                 |
| ADGRA2      | 5     | -      | adhesion G protein-coupled receptor A2                    |
| IFFO1       | 5     | -      | intermediate filament family orphan 1                    |
| CHTF18      | 5     | -      | chromosome transmission fidelity factor 18                |
| TMEM204     | 5     | -      | transmembrane protein 204                                |
| RECQL5      | 5     | -      | RecQ like helicase 5                                     |
| SMIM5       | 5     | -      | small integral membrane protein 5                        |
| MAPK1       | 5     | PK     | mitogen-activated protein kinase 1                        |
| SYN1        | 5     | -      | synapsin 1                                                |

Table 9: For genes with at least 5 hemimethylation sites in tumor samples. The gene name, corresponding number of hemimethylated sites, specified gene family, and a description of the gene are formatted in the table's respective columns. Descriptions are derived from the Molecular Signature Database [25] and the GeneCards database [26]. Certain genes are indicated as members of specific gene families, as shown in the third column: “TF” for transcription factor, “TS” for tumor suppressor, “OG” for oncogene, “HP” for homeodomain protein, “TCG” for translocated cancer gene, and “PK” for protein kinase.
| Gene name | Count | Family | Gene Description |
|-----------|-------|--------|------------------|
| ZFPM1     | 14    | TF     | zinc finger protein, FOG family member 1 |
| GNAS      | 13    | OG     | GNAS complex locus |
| RGPD2     | 12    | -      | RANBP2 like and GRIP domain containing 2 |
| SHANK3    | 11    | -      | SH3 and multiple ankyrin repeat domains 3 |
| IRX2      | 10    | TF, HP | iroquois homeobox 2 |
| LTB4R     | 9     | -      | leukotriene B4 receptor |
| CPEB3     | 8     | -      | cytoplasmic polyadenylation element binding protein 3 |
| PTPRN2    | 7     | -      | protein tyrosine phosphatase receptor type N2 |
| MIR1268A  | 7     | -      | microRNA 1268a |
| GNAS-AS1  | 7     | -      | GNAS antisense RNA 1 |
| CYP26C1   | 7     | -      | cytochrome P450 family 26 subfamily C member 1 |
| TBL1XR1   | 6     | -      | transducin beta like 1 X-linked receptor 1 |
| HOXA3     | 6     | TF, HP | homeobox A3 |
| CACNA1H   | 6     | -      | calcium voltage-gated channel subunit alpha1 H |
| NPEPPS    | 6     | -      | aminopeptidase puromycin sensitive |
| SEMA6B    | 6     | CGF    | semaphorin 6B |
| HOMER3    | 6     | -      | homer scaffold protein 3 |
| PINLYP    | 6     | -      | phospholipase A2 inhibitor and LY6/PLAUR domain containing |
| GDI1      | 6     | -      | GDP dissociation inhibitor 1 |
| HS3ST2    | 6     | -      | heparan sulfate-glucosamine 3-sulfotransferase 2 |
| PRDM16    | 5     | TF, OG, TCG | PR/SET domain 16 |
| PLK3      | 5     | PK     | polo like kinase 3 |
| GREM2     | 5     | CGF    | gremlin 2, DAN family BMP antagonist |
| MEIS1     | 5     | TF, HP | Meis homeobox 1 |
| MEIS1-AS2 | 5     | -      | MEIS1 antisense RNA 2 |
| POLH      | 5     | -      | DNA polymerase eta |
| HOXA-AS2  | 5     | -      | HOXA cluster antisense RNA 2 |
| EBF3      | 5     | -      | EBF transcription factor 3 |
| CBFA2T3   | 5     | TF, OG, TCG | CBFA2/RUNX1 translocation partner 3 |
| RPL13     | 5     | -      | ribosomal protein L13 |
| NFIC      | 5     | TF     | nuclear factor 1C |
| CDH4      | 5     | -      | cadherin 4 |
| PDGFB     | 5     | OG, TCG | cytokine or growth factor, platelet derived growth factor subunit B |
| CCNT1     | 5     | -      | cyclin T1 |
| SNORD68   | 5     | -      | small nucleolar RNA, C/D box 68 |

Table 10: For genes with at least 5 hemimethylation sites in normal samples. The gene name, corresponding number of hemimethylated sites, specified gene family, and a description of the gene are formatted in the table’s respective columns. Descriptions are derived from the Molecular Signature Database [25] and the GeneCards database [26]. Certain genes are indicated as members of specific gene families, as shown in the third column: “TF” for transcription factor, “TS” for tumor suppressor, “OG” for oncogene, “HP” for homeodomain protein, “TCG” for translocated cancer gene, and “PK” for protein kinase.
| Gene name  | Count | Family | Gene Description                                                                 |
|------------|-------|--------|-----------------------------------------------------------------------------------|
| RGPD5      | 16    | -      | RANBP2 like and GRIP domain containing 5                                            |
| RGPD8      | 16    | -      | RANBP2 like and GRIP domain containing 8                                             |
| ROCK1P1    | 13    | -      | Rho associated coiled-coil containing protein kinase 1 pseudogene 1                 |
| THAP4      | 8     | -      | THAP domain containing 4                                                            |
| SGT2A      | 8     | -      | small glutamine rich tetratricopeptide repeat containing alpha                       |
| PTPRN2     | 7     | -      | protein tyrosine phosphatase receptor type N2                                        |
| CNTNAP3    | 7     | -      | contactin associated protein like 3                                                 |
| NUTM2A-AS1 | 7     | -      | NUTM2A antisense RNA 1                                                              |
| RBFOX3     | 7     | -      | RNA binding fox-1 homolog 3                                                         |
| ESPNP      | 6     | -      | espin pseudogene                                                                   |
| FOXK1      | 6     | -      | forkhead box K1                                                                    |
| H0XA3      | 6     | HP, TF | homeobox A3                                                                        |
| LMF1       | 6     | -      | lipase maturation factor 1                                                          |
| USP45      | 6     | -      | ubiquitin specific peptidase 45                                                     |
| LOC101928782 | 6 | -      | RNA Gene affiliated with the lncRNA class                                           |
| PRDM16     | 5     | OG, TF, TCG | PR/SET domain 16                                                                 |
| RGPD4      | 5     | -      | RANBP2 like and GRIP domain containing 4                                             |
| MERTK      | 5     | PK     | MER proto-oncogene, tyrosine kinase                                                  |
| FAM160A1   | 5     | -      | family with sequence similarity 160 member A1                                        |
| PRKAR1B    | 5     | -      | protein kinase cAMP-dependent type 1 regulatory subunit beta                        |
| MAD1L1     | 5     | -      | mitotic arrest deficient 1 like 1                                                    |
| H0XA2      | 5     | HP, TF | homeobox A2                                                                        |
| DPP6       | 5     | -      | dipeptidyl peptidase like 6                                                         |
| DIP2C      | 5     | -      | disco interacting protein 2 homolog C                                               |
| FANK1      | 5     | -      | fibronectin type III and ankyrin repeat domains 1                                    |
| GAL3ST3    | 5     | -      | galactose-3-O-sulfotransferase 3                                                    |
| FLJ12825   | 5     | -      | RNA Gene affiliated with the lncRNA class                                            |
| KLF5       | 5     | TF     | Kruppel like factor 5                                                               |
| ISL2       | 5     | HP, TF | ISL LIM homeobox 2                                                                  |
| CBFA2T3    | 5     | OG, TF, TCG | CBFA2/RUNX1 translocation partner 3                                                |
| SBN02      | 5     | -      | strawberry notch homolog 2                                                          |
| GIPR       | 5     | -      | gastric inhibitory polypeptide receptor                                              |
| SCAF1      | 5     | -      | SR-related CTD associated factor 1                                                   |
| COL6A1     | 5     | -      | collagen type VI alpha 1 chain                                                       |
| NEXMIF     | 5     | -      | neurite extension and migration factor                                               |
| GK5        | 5     | -      | glycerol kinase 5                                                                   |

**Table 11:** For genes with at least 5 hemimethylation sites in both tumor and normal samples. The gene name, corresponding number of hemimethylated sites, specified gene family, and a description of the gene are formatted in the table’s respective columns. Descriptions are derived from the Molecular Signature Database [25] and the GeneCards database [26]. Certain genes are indicated as members of specific gene families, as shown in the third column: “TF” for transcription factor, “TS” for tumor suppressor, “OG” for oncogene, “HP” for homeodomain protein, “TCG” for translocated cancer gene, and “PK” for protein kinase.

In order to understand the functions and relationships of these genes, we further analyze their biological interactions using the ConsensusPath Database (CPDB) software package [27-29], see Figures 4, 5, 6, and 7. Figure 4 describes the different types of biological relationships between genes based on the CPDB software. A gene with a black label is known to be hemimethylated (i.e., identified by our analysis).
and a gene with a purple label is a gene that is not provided in our hemimethylation gene list but it interacts with one of the known genes. Figure 4 is the legend for Figures 5, 6, and 7. This legend figure summarizes the relationships for gene lists in Tables 9, 10, and 11 as shown in Figures 5, 6, and 7, respectively. These figures show the extent to which these highly hemimethylated genes interact and possibly affect the cell function of related genes.

Figure 5 shows genetic interactions between genes with the most hemimethylation in tumor samples, and these genes are recorded in Table 9. The gene network in Figure 5 contains a number of hub genes with complex interactions. These hub genes include GNAS, NFATC1, NOTCH1, MAPK1, HOAC4, TP73, and EGR1. We can see that if a hub gene like MAPK1 is hemimethylated, it may interact with dozens of other genes. Some of these genes, e.g., EGR1 [30-33] and UNC5B [34-37], are known to be associated with different cancers, including lung cancer. EGR1 has a promoting effect on cancer metastasis in OCT4-overexpressing lung cancer [38]. The pseudogene DUXAP8 may act as an oncogene in non-small-cell lung cancer, and it may play this role by silencing EGR1 and RHOB transcription via binding with EZH2 and LSD1 [39]. The expression of UNC5A, UNC5B, or UNC5C is down-regulated in multiple cancers including lung cancer [40], and UNC5B has also been indicated as a putative tumor suppressor [41].

Figure 6 shows genetic interactions between genes with the most hemimethylation in normal DNA, and these genes are recorded in Table 10. In this figure, we can see that GNAS is a hub gene interacting with many other genes that may not be hemimethylated themselves. GNAS is observed in both tumor and normal samples, as well as in the hemimethylation study for breast cancer cell lines [9]. MEIS1 is also a hub gene that interacts with genes like KMT2A [42] and TK1 [43]. While these genes are not hemimethylated in our samples, they are known to be associated with cancer. KMT2A and hTERT are positively correlated in melanoma tumor tissues, and KMT2A promotes melanoma cell growth by targeting the hTERT signaling pathway [44]. KMT2A has an epigenetic regulation role on NOTCH1 and NOTCH3, and this mechanism is essential for inhibiting glioma proliferation [45]. TK1 plays a moderate role as a diagnostic tumor marker for cancer patients [46], and it is a potential clinical biomarker for the treatment of lung, breast, and colorectal cancer [47]. A systematic review shows that TK1 overexpression is associated with the poor outcomes of lung cancer patients [48]. MEIS1 inhibits non-small cell lung cancer cell proliferation [49]. MEIS1 plays a crucial role in normal development [26] and it is also reported as an important gene related to leukemia [50-52]. Therefore, it is possible that the hemimethylation of hub genes like MEIS1 affects protein, biochemical, or regulatory functions of genes that are associated with cancer.

Figure 7 shows genetic interactions between genes with the most hemimethylation on identical locations in tumor and normal samples. These genes are recorded in Table 11. This means that the hemimethylation of CpG sites in this network is unchanged or unaffected by the formation of cancer. The HNRNPL gene is a major hub in this gene network. While we do not detect any hemimethylation in this gene, it directly interacts with 10 genes that we know to be hemimethylated. Some of these genes, like PTPRN2 and MAD1L1, can also be found in the tumor gene network, see Figure 5. There appears to be no common genes between Figure 6 (hemimethylated genes in normal samples) and Figure 7.
(hemimethylated genes in both tumor and normal samples). Therefore, genes that have a large number of hemimethylated CpG sites found only in normal DNA seem to have very few CpG sites that remain the same when cancer forms.

In addition to the above analysis, we have conducted gene set enrichment analysis using the molecular signature database and the related software package provided by the Broad Institute [25]. Of the most hemimethylated genes in tumor DNA (Table 9), six are also significantly represented in cancer module 163 (with p-value < 0.05). This module is a collection of genes known to be overrepresented in cancer pathways and is reported by the Stanford research group [53]. The six genes are IFT140, IFFO1, SYN1, FMNL1, NOTCH1, and RGS14. There are no such overly represented genes and cancer modules among genes shown in Table 10 (for normal samples) and Table 11 (for both tumor and normal samples).

**Discussion**

The p-value used in these results is 0.05 and the mean difference cutoff values, 0.4, 0.6 and 0.8, are predetermined based on our previous research [9]. Results are narrowed down to the 0.4 cutoff level to allow more results to be viewed, as the higher cutoff values restricted the available hemimethylated CpG sites from being identified. The number of both tumor and normal clusters detected decreases rapidly as we increase the mean cutoff value at each CpG site as shown in Table 2. With more strict criteria, the methylation difference between the two DNA strands at each CpG site must exist in order for us to consider hemimethylation at a CpG site. This rapid decrease may indicate certain hemimethylation heterogeneity in lung cancer as cancer methylation patterns are generally heterogeneous among multiple patients or cell lines [54].

For the 41 most hemimethylated genes in lung cancer tumors, 7 of them are also highly hemimethylated in breast cancer cell lines, as reported by Sun et al. [9]. These seven genes are PRDM16, GNAS, PTPRN2, MAD1L1, HDAC4, NOTCH1, and CACNA1H. The remaining 34 highly hemimethylated genes in the lung tumor sample are not highly hemimethylated in breast cancer cell lines. It is possible that these genes are unique to lung cancer and as a result would be helpful when diagnosing patients with lung cancer specifically, but further research needs to be done.

Based upon the outcome of hemimethylation research in breast cancer cell lines, the frequency of polarity clusters is much higher than the one in this paper. The results of breast cancer hemimethylation analysis indicate polarity clusters are more frequently found than regular clusters [9]. However, the lung cancer analysis reflects contrasting results; polarity clusters are less frequently found than regular clusters for both tumor and normal samples, as shown in Figure 2, Table 4, and Table 5. The source of data may have some influence on this. The breast cancer study is performed using breast cancer cell lines, which are tumors grown in labs over a long period of time. Whereas, this study uses primary tissues directly from lung cancer patients. Another factor could be the type of cancer, as hemimethylation pattern frequency may be tissue specific. Future research into the biological impact of hemimethylation on different kinds of cells may allow for more insight into these differences and their effects.
After conducting statistical tests for a large number of CpG sites, selecting the significant CpG sites is a crucial step, and multiple testing correction is important because using only the raw p-values may result in many false positive sites. However, for the understudied hemimethylation pattern, the proper way of doing multiple testing correction is not clear. In order to explore the impact of different correction methods, we have used the following three methods: a simple moving-average based method, the comb-p FDR method, and the comb-p SLK method. Note, comb-p is a software package developed for combining, analyzing, and correcting spatially correlated p-values [55]. FDR stands for the Benjamini–Hochberg false discovery correction [56]. SLK represents the Stouffer–Liptak–Keohrs correction [57].

For the simple moving-average based p-value correction, the p-value for each CpG site is averaged with the p-values of its two neighboring sites within d (d=25, 50, and 100) base pairs on each site of the genome. If either of the neighbors is further than the d base pairs away, they are not counted in the average. The results in the top panel (top 4 columns) of Table 12 show that adjusting the p-values using moving average results in a large decrease in the number of significant CpG sites. However, it still identifies a much larger number of CpG sites than the number of sites selected using both mean difference and p-value as shown in the Table 12 bottom panel. The differences are 36,521 vs 7,351 (for normal) and 33,431 vs. 7,330 (for tumor). When we apply the mean difference filter after adjusting the p-values with the moving average method, the number of significant CpG sites is further reduced to 3,354 in normal and 3,030 in tumor samples. It is a very strict multiple testing correction method when using the moving average combined with the mean difference. This should be avoided because overly corrected results may introduce false negatives and disregard real hemimethylated sites, which can be rare and hard to identify.

|               | p<0.05 | p<0.05.d=25 | p<0.05.d=50 | p<0.05.d=100 |
|---------------|--------|-------------|-------------|--------------|
| Normal        | 102751 | 36521       | 26399       | 23641        |
| Tumor         | 89326  | 33431       | 23471       | 20714        |

| MeanDiff MeanDiff | p<0.05 | MeanDiff p<0.05.d=25 | MeanDiff p<0.05.d=50 | MeanDiff p<0.05.d=100 |
|-------------------|--------|----------------------|----------------------|-----------------------|
| Normal            | 7351   | 3354                 | 2572                 | 2329                  |
| Tumor             | 7330   | 3030                 | 2303                 | 2056                  |

**Table 12**: Number of significant CpG sites selected based on different criteria. The three rows in the top panel are the CpG sites selected based on p-value only; the last three rows in the bottom panel are the CpG sites selected based on both mean difference (MeanDiff) and p-value. The d=25, d=50, and d=100 column names correspond to the distance (in base pairs) that we used to calculate the moving-average of neighboring p-values. For example, “p<0.05.d=25” means the new p-value is obtained using the moving average of nearby CpG sites within d=25 base pairs on each side, and then we use p<0.05 to select significant CpG sites.

The results of using comb-p FDR and SLK multiple testing corrections are shown in Table 13. To simplify our analysis, we only show the results of analyzing chromosome 1 (chr1) CpG sites. Table 13 shows that
the comb-p FDR correction is too strict because we have no significant hemimethylation sites identified at the 0.05 cutoff. The comb-p SLK correction can add more false negatives. For example, for the chr1 normal data, if we only use the raw p value<0.05, we identify 8,757 sites. If we use comb-p SLK, we identify 5,916 sites. Among these two lists of significant sites, there are only 2,602 sites identified by both selection methods. Therefore, using the comb-p SLK method may lead to the identification of 3,314 (5,916 – 2,602) false positive CpG sites for which the original data did not show a small p-value. This seems counterintuitive, but Table 14 explains why. In Table 14, we run the comb-p based on the SLK correction (with default settings) for an example dataset of 100 sites provided in the comb-p web page. The table contains the top 10 rows and shows us that the SLK method identifies significant regions and then assigns all CpG sites in one region the same p-value. The SLK method may overcorrect the results and introduce both false negative and false positive sites. For example, Table 14 shows that the site with the start position 62,140 is not significant (with p-value = 0.79429). However, after the SLK correction, the modified p-value is 8.80E-05, which is a false-positive result.

| chr1 | p<0.05 | comb-p.FDR.p<0.05 | comb-p.SLK | SLK.Intersect |
|------|--------|-------------------|------------|---------------|
| Normal | 8757 | 0 | 5916 | 2602 |
| Tumor | 7209 | 0 | 4275 | 2155 |

Table 13: The number of significant CpG sites selected for chromosome 1 (chr1). The “p<0.05” column describes the number of significant sites identified in chromosome 1. The “comb-p.FDR.p<0.05” column shows the number of significant sites after FDR correction. The “comb-p.SLK” column describes the number of significant sites after SLK correction. Lastly, the “SLK.Intersect” column shows the number of CpG sites that are significant both before and after the SLK correction.

| chr | start | end | p-value | SLK.p-value |
|----|-------|-----|---------|-------------|
| chr1 | 62082 | 62083 | 3.24E-08 | 8.80E-05 |
| chr1 | 62096 | 62097 | 3.94E-12 | 8.80E-05 |
| chr1 | 62112 | 62113 | 0.064892 | 8.80E-05 |
| chr1 | 62140 | 62141 | 0.79429 | 8.80E-05 |
| chr1 | 526852 | 526853 | 0.45423 | 0.50846 |
| chr1 | 526880 | 526881 | 0.13813 | 0.50846 |
| chr1 | 526889 | 526890 | 0.45877 | 0.50846 |
| chr1 | 526896 | 526897 | 0.73267 | 0.50846 |
| chr1 | 526910 | 526911 | 0.53226 | 0.50846 |
| chr1 | 526930 | 526931 | 0.58344 | 0.50846 |

Table 14: An example of the multiple testing correction p-value based on the SLK method. Each row corresponds to one CpG site. The first, second, and third columns describe the location of each CpG site. The “p-value” column gives the raw p-values with no correction. The last column, “SLK.p-value”, describes the p-values after SLK correction. After exploring various correction methods, we conclude that the mean difference plus p-value filtering method can produce meaningful and interpretable results when dealing with the multiple testing
comparison problem for our hemimethylation analysis. As shown in Table 12, applying the mean difference cutoff alone eliminates over 90% of p-values (e.g., 102,751 vs 7,351 for the normal, and 89,326 vs 7,330 for the tumor). In this context, a false positive means that the p-value suggests a statistical significance but there is not a practical (or biological) difference. False positive sites are unlikely to have large mean differences due to there being no significant biological difference. The false positive sites would be eliminated at a higher rate by the mean difference cutoff. Therefore, applying any additional corrections would be too strict and result in more false negatives. We have finally decided to use the p-value and mean difference to identify significant hemimethylation sites.

As for the criteria we used to identify hemimethylation sites, in addition to p-values and mean differences, we may add an additional one. That is, the methylation signal (or the count of methylated reads) on one DNA strand is zero, and the methylation signal (or the count of reads) on the other strand is positive. Adding this criteria will help us to identify hemimethylation sites more strictly as no methylation on one strand and positive (or high) methylation signal on another strand. This criterion is ideal when the datasets are not very noisy and when the samples are homogeneous (i.e., not heterogeneous). In this project, we choose to only use p-value and mean difference for the following reasons. First, we use these two criteria to make a fair comparison between the previous breast cancer results [9] and our new lung cancer results. Second, the average methylation signals at most CpG sites tend to be clustered around 0 or 1 [58]. Third, bisulfite converted methylation sequencing data can be noisy and tumor samples’ methylation signals are very heterogeneous. We generally consider that even though the average methylation signals around 0 to 0.2 or 0.25 are still roughly in the category of no or very low methylation signals [54]. Therefore, most of the CpG sites we identified still tend to have relatively high methylation signals on one DNA strand and have relatively low or no methylation signals on another strand.

**Conclusion**

Hemimethylation is an important but understudied pattern in cancer. In this paper, we have conducted the first-ever exploratory investigation of hemimethylation in lung cancer. In particular, we have conducted statistical analyses to identify hemimethylation patterns for non-small cell lung cancer patients. We have identified both singleton hemimethylation sites and different clusters in normal and tumor cells. We have also conducted bioinformatic analysis on the genes that have relatively more hemimethylated sites in tumor, normal, and both tumor and normal cells to see the biological interactions of these genes. Our results show that not only does hemimethylation exist in lung cells, but also with diverse patterns and frequencies that are comparable between normal and tumorous cells. We conclude that hemimethylation is related to both normal and tumor cell development. This is also seen by its existence in the same genes in normal and lung tumor cells. However, there are certain genes that only have hemimethylated sites for one type of cell, normal or tumor, but not both. Certain genes are previously known to be associated with carcinogenesis. These genes exhibit existence in one cell type and not the other. Hemimethylation existing in this way may imply epigenetic changes in certain genes associated with lung cancer. The development and progression of lung cancer may be tracked by the analysis of epigenetic change (i.e., hemimethylation and methylation) in these regions.
Declarations

Ethics approval and consent to participate

No ethics approval is required for the study.

Consent for publication

Not applicable.

Availability of data and materials

Datasets analyzed for this study are publicly available (SRP125064) and can be downloaded from this web page: https://www.ncbi.nlm.nih.gov/sra/SRP125064. R code files are available upon request.

Competing interests

The authors declare that they have no competing interests.

Funding

This research was partially supported by the NSF-REU grant DMS-1757233 (June – August, 2019). This project is also supported by the Texas State University Research Enhancement Program (an internal award for Dr. Sun). The funders did not play any role in the study design, data analysis, interpretation of data, writing the manuscript, or decision to publish.

Authors’ contributions

SS initiated the project, suggested all key original ideas, and led the whole process. AZ contributed to most of the statistical analysis and the revision of this paper. CF helped with the data analysis and the paper revision. JP did minor coding work in the beginning. AZ and CF contributed significantly to the writing and editing of the manuscript. SS gave suggestions over the course of the project and extensively reviewed and revised the final paper. All authors contributed expertise and edits. All authors have read and approved the final manuscript.

Acknowledgements

We appreciate the help and support of Texas State University Writing Center. We are grateful for the two reviewers’ thoughtful comments.

Abbreviations

CpG: The shorthand notation for 5’-cytosine-phosphate-guanine-3’.
MU: When it is for one CpG site on two DNA strands, it refers to a hemimethylated CpG site with methylation (M) on the positive strand and unmethylation (U) on the reverse strand. When it refers to two consecutive CpG sites on one DNA strand, it means that methylation occurs on the first CpG site (i.e., M), but not on the second one (i.e., U).

UM: When it is for one CpG site on two DNA strands, it refers to a hemimethylated CpG site with unmethylation (U) on the positive strand and methylation (M) on the reverse strand. When it refers to two consecutive CpG sites on one DNA strand, it means that methylation does not occur on the first CpG site (i.e., U), but occurs on the second one (i.e., M).

NS: A CpG site identified as not significantly (NS) hemimethylated.

References

1. American Cancer Society (www.cancer.org).
2. Howlader N, Noone AM, Krapcho M, Miller D, Brest A, Yu M, Ruhl J, Tatalovich Z, Mariotto A, Lewis DR et al: SEER Cancer Statistics Review, 1975-2016, National Cancer Institute. Bethesda, MD, https://seer.cancer.gov/csr/1975_2016/, based on November 2018 SEER data submission, posted to the SEER web site, April 2019. 2019.
3. Lim DH, Maher E: DNA methylation: a form of epigenetic control of gene expression. The Obstetrician & Gynaecologist 2010, 12:6.
4. Sharif J, Koseki H: Hemimethylation: DNA's lasting odd couple. Science 2018, 359(6380):1102-1103.
5. Ehrlich M, Lacey M: DNA hypomethylation and hemimethylation in cancer. Advances in Experimental Medicine and Biology 2013, 754:31-56.
6. Li E, Zhang Y: DNA methylation in mammals. Cold Spring Harb Perspect Biol 2014, 6(5):a019133.
7. Shao C, Lacey M, Dubéau L, Ehrlich M: Hemimethylation footprints of DNA demethylation in cancer. Epigenetics 2009, 4(3):165-175.
8. Xu C, Corces VG: Nascent DNA methylome mapping reveals inheritance of hemimethylation at CTCF/cohesin sites. Science 2018, 359(6380):1166-1170.
9. Sun S, Lee YR, Enfield B: Hemimethylation Patterns in Breast Cancer Cell Lines. Cancer Inform 2019, 18:1176935119872959.
10. Sun X, Han Y, Zhou L, Chen E, Lu B, Liu Y, Pan X, Cowley AW, Jr., Liang M, Wu Q et al: A comprehensive evaluation of alignment software for reduced representation bisulfite sequencing data. Bioinformatics 2018, 34(16):2715-2723.
11. Harris EY, Ponts N, Le Roch KG, Lonardi S: BRAT-BW: efficient and accurate mapping of bisulfite-treated reads. Bioinformatics 2012, 28(13):1795-1796.
12. Sun S, Li P: HMPL: A Pipeline for Identifying Hemimethylation Patterns by Comparing Two Samples. Cancer Inform 2015, 14(Suppl 2):235-245.
13. Wilcoxon F: Individual Comparisons by Ranking Methods. Biometrics Bulletin 1945, 1(6):80-83.
14. Doherty R, Couldrey C: Exploring genome wide bisulfite sequencing for DNA methylation analysis in livestock: a technical assessment. Front Genet 2014, 5:126.

15. Rodriguez N, Pelaez A, Barderas R, Dominguez G: Clinical implications of the deregulated TP73 isoforms expression in cancer. Clin Transl Oncol 2018, 20(7):827-836.

16. Yao Z, Di Poto C, Mavodza G, Oliver E, Ressom HW, Sherif ZA: DNA Methylation Activates TP73 Expression in Hepatocellular Carcinoma and Gastrointestinal Cancer. Sci Rep 2019, 9(1):19367.

17. Ye H, Guo X: TP73 is a credible biomarker for predicting clinical progression and prognosis in cervical cancer patients. Biosci Rep 2019, 39(8).

18. Hollstein PE, Shaw RJ: GNAS shifts metabolism in pancreatic cancer. Nat Cell Biol 2018, 20(7):740-741.

19. Idziaszczyk S, Wilson CH, Smith CG, Adams DJ, Cheadle JP: Analysis of the frequency of GNAS codon 201 mutations in advanced colorectal cancer. Cancer Genet Cytogenet 2010, 202(1):67-69.

20. Ikuta K, Seno H, Chiba T: Molecular changes leading to gastric cancer: a suggestion from rare-type gastric tumors with GNAS mutations. Gastroenterology 2014, 146(5):1417-1418.

21. Jin X, Zhu L, Cui Z, Tang J, Xie M, Ren G: Elevated expression of GNAS promotes breast cancer cell proliferation and migration via the PI3K/AKT/Snail1/E-cadherin axis. Clin Transl Oncol 2019, 21(9):1207-1219.

22. Tominaga E, Tsuda H, Arao T, Nishimura S, Takano M, Kato H, Minami A, Aoki D, Nishio K: Amplification of GNAS may be an independent, qualitative, and reproducible biomarker to predict progression-free survival in epithelial ovarian cancer. Gynecologic oncology 2010, 118(2):160-166.

23. Gan RH, Wei H, Xie J, Zheng DP, Luo EL, Huang XY, Xie J, Zhao Y, Ding LC, Su BH et al: Notch1 regulates tongue cancer cells proliferation, apoptosis and invasion. Cell Cycle 2018, 17(2):216-224.

24. Zeng JS, Zhang ZD, Pei L, Bai ZZ, Yang Y, Yang H, Tian QH: CBX4 exhibits oncogenic activities in breast cancer via Notch1 signaling. Int J Biochem Cell Biol 2018, 95:1-8.

25. Subramanian A, Tamayo P, Mootha VK, Mukherjee S, Ebert BL, Gillette MA, Paulovich A, Pomeroy SL, Golub TR, Lander ES et al: Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. Proc Natl Acad Sci USA 2005, 102(43):15545-15550.

26. GeneCards - Gene Database (www.genecards.org)

27. Kamburov A, Pentchev K, Galicka H, Wierling C, Lehrach H, Herwig R: ConsensusPathDB: toward a more complete picture of cell biology. Nucleic acids research 2011, 39(Database issue):D712-717.

28. Kamburov A, Stelzl U, Lehrach H, Herwig R: The ConsensusPathDB interaction database: 2013 update. Nucleic acids research 2013, 41(Database issue):D793-800.

29. Kamburov A, Wierling C, Lehrach H, Herwig R: ConsensusPathDB—a database for integrating human functional interaction networks. Nucleic acids research 2009, 37(Database issue):D623-628.

30. Redmond KL, Crawford NT, Farmer H, D’Costa ZC, O’Brien GJ, Buckley NE, Kennedy RD, Johnston PG, Harkin DP, Mullan PB: T-box 2 represses NDRG1 through an EGR1-dependent mechanism to drive the
proliferation of breast cancer cells. *Oncogene* 2010, **29**(22):3252-3262.

31. Shajahan-Haq AN, Boca SM, Jin L, Bhuvaneshwar K, Gusev Y, Cheema AK, Demas DD, Raghavan KS, Michalek R, Madhavan S *et al*.: *EGR1 regulates cellular metabolism and survival in endocrine resistant breast cancer*. *Oncotarget* 2017, **8**(57):96865-96884.

32. Wenzel K, Daskalow K, Herse F, Seitz S, Zacharias U, Schenk JA, Schulz H, Hubner N, Micheel B, Schlag PM *et al*.: *Expression of the protein phosphatase 1 inhibitor KEPI is downregulated in breast cancer cell lines and tissues and involved in the regulation of the tumor suppressor EGR1 via the MEK-ERK pathway*. *Biol Chem* 2007, **388**(5):489-495.

33. Yang M, Teng W, Qu Y, Wang H, Yuan Q: *Sulforaphene inhibits triple negative breast cancer through activating tumor suppressor Egr1*. *Breast Cancer Res Treat* 2016, **158**(2):277-286.

34. Kong C, Zhan B, Piao C, Zhang Z, Zhu Y, Li Q: *Overexpression of UNC5B in bladder cancer cells inhibits proliferation and reduces the volume of transplantation tumors in nude mice*. *BMC Cancer* 2016, **16**(1):892.

35. Liu J, Kong CZ: [Expressions of netrin-1 and UNC5B in prostate cancer and their clinical significance]. *Zhonghua Nan Ke Xue* 2013, **19**(12):1072-1076.

36. Liu J, Zhang Z, Li ZH, Kong CZ: *Clinical significance of UNC5B expression in bladder cancer*. *Tumour Biol* 2013, **34**(4):2099-2108.

37. Okazaki S, Ishikawa T, Iida S, Ishiguro M, Kobayashi H, Higuchi T, Enomoto M, Mogushi K, Mizushima H, Tanaka H *et al*.: *Clinical significance of UNC5B expression in colorectal cancer*. *Int J Oncol* 2012, **40**(1):209-216.

38. Feng YH, Su YC, Lin SF, Lin PR, Wu CL, Tung CL, Li CF, Shieh GS, Shiau AL: *Oct4 upregulates osteopontin via Egr1 and is associated with poor outcome in human lung cancer*. *BMC Cancer* 2019, **19**(1):791.

39. Sun M, Nie FQ, Zang C, Wang Y, Hou J, Wei C, Li W, He X, Lu KH: *The Pseudogene DUXAP8 Promotes Non-small-cell Lung Cancer Cell Proliferation and Invasion by Epigenetically Silencing EGR1 and RHOB*. *Mol Ther* 2017, **25**(3):739-751.

40. Thiebault K, Mazelin L, Pays L, Llambi F, Joly MO, Scoazec JY, Saurin JC, Romeo G, Mehlen P: *The netrin-1 receptors UNC5H are putative tumor suppressors controlling cell death commitment*. *Proc Natl Acad Sci U S A* 2003, **100**(7):4173-4178.

41. Xu H, Han Y, Liu B, Li R: *Unc-5 homolog B (UNC5B) is one of the key downstream targets of N-α-Acetyltransferase 10 (Naa10)*. *Scientific Reports* 2016, **6**.

42. Rao RC, Dou Y: *Hijacked in cancer: the KMT2 (MLL) family of methyltransferases*. *Nat Rev Cancer* 2015, **15**(6):334-346.

43. Bagegni N, Thomas S, Liu N, Luo J, Hoog J, Northfelt DW, Goetz MP, Forero A, Bergqvist M, Karen J *et al*.: *Serum thymidine kinase 1 activity as a pharmacodynamic marker of cyclin-dependent kinase 4/6 inhibition in patients with early-stage breast cancer receiving neoadjuvant palbociclib*. *Breast Cancer Res* 2017, **19**(1):123.
44. Zhang C, Song C, Liu T, Tang R, Chen M, Gao F, Xiao B, Qin G, Shi F, Li W et al: **KMT2A promotes melanoma cell growth by targeting hTERT signaling pathway.** *Cell Death Dis* 2017, 8(7):e2940.

45. Huang YC, Lin SJ, Shih HY, Chou CH, Chu HH, Chiu CC, Yuh CH, Yeh TH, Cheng YC: **Epigenetic regulation of NOTCH1 and NOTCH3 by KMT2A inhibits glioma proliferation.** *Oncotarget* 2017, 8(38):63110-63120.

46. Xiang Y, Zeng H, Liu X, Zhou H, Luo L, Duan C, Luo X, Yan H: **Thymidine kinase 1 as a diagnostic tumor marker is of moderate value in cancer patients: A meta-analysis.** *Biomed Rep* 2013, 1(4):629-637.

47. Weagel EG, Burrup W, Kovtun R, Velazquez EJ, Felsted AM, Townsend MH, Ence ZE, Suh E, Piccolo SR, Weber KS et al: **Membrane expression of thymidine kinase 1 and potential clinical relevance in lung, breast, and colorectal malignancies.** *Cancer Cell Int* 2018, 18:135.

48. Wei YT, Luo YZ, Feng ZQ, Huang QX, Mo AS, Mo SX: **TK1 overexpression is associated with the poor outcomes of lung cancer patients: a systematic review and meta-analysis.** *Biomark Med* 2018, 12(4):403-413.

49. Li W, Huang K, Guo H, Cui G: **Meis1 regulates proliferation of non-small-cell lung cancer cells.** *J Thorac Dis* 2014, 6(6):850-855.

50. Li Z, Huang H, Chen P, He M, Li Y, Arnovitz S, Jiang X, He C, Hyjek E, Zhang J et al: **Publisher Correction: miR-196b directly targets both HOXA9/MEIS1 oncogenes and FAS tumour suppressor in MLL-rearranged leukaemia.** *Nat Commun* 2018, 9:16192.

51. Li Z, Huang H, Chen P, He M, Li Y, Arnovitz S, Jiang X, He C, Hyjek E, Zhang J et al: **miR-196b directly targets both HOXA9/MEIS1 oncogenes and FAS tumour suppressor in MLL-rearranged leukaemia.** *Nat Commun* 2012, 3:688.

52. Imamura T, Morimoto A, Takanashi M, Hibi S, Sugimoto T, Ishii E, Imashuku S: **Frequent co-expression of HoxA9 and Meis1 genes in infant acute lymphoblastic leukaemia with MLL rearrangement.** *Br J Haematol* 2002, 119(1):119-121.

53. **http://robotics.stanford.edu/~erans/cancer/modules/** .

54. Tian S, Bertelsmann K, Yu L, Sun S: **DNA Methylation Heterogeneity Patterns in Breast Cancer Cell Lines.** *Cancer Inform* 2016, 15(Supple 4):1-9.

55. Pedersen BS, Schwartz DA, Yang IV, Kechrins KJ: **Comb-p: software for combining, analyzing, grouping and correcting spatially correlated P-values.** *Bioinformatics* 2012, 28(22):2986-2988.

56. Benjamini Y, Hochberg Y: **Controlling the false discovery rate: a practical and powerful approach to multiple testing.** *Journal of the Royal Statistical Society* 1995, 57(1):289-300.

57. Kechrins KJ, Biehs B, Kornberg TB: **Generalizing moving averages for tiling arrays using combined p-value statistics.** *Stat Appl Genet Mol Biol* 2010, 9:Article29.

58. Xu L, Mitra-Behera S, Alston B, Zong Z, Sun S: **Identifying DNA methylation variation patterns to obtain potential breast cancer biomarker genes.** *International Journal of Biomedical Data Mining* 2015, 4(1):1-8.