Statistical and Bioinformatic Analysis of Hemimethylation Patterns in 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

Posted Date: March 20th, 2020

DOI: https://doi.org/10.21203/rs.3.rs-17794/v1

Version of Record: A version of this preprint was published on March 12th, 2021. See the published version at https://doi.org/10.1186/s12885-021-07990-7.

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 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 impact. 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 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 in-depth investigation of hemimethylation for lung cancer. Our results show that hemimethylation does exist in lung cells either as singletons or clusters. Most clusters contain only 2 or 3 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 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 contribute to the development of hemimethylation, which means that methylation occurs on only one DNA strand of a CpG site and not the other, see Fig. 1. The exact process by which the methyl groups are lost or gained is not well understood but may one day provide new insight into the development of cancerous cells [5].

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 an "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 in consecutive ones, known as hemimethylation clusters. Such clusters can manifest in one of two distinct patterns: regular or polar [5, 7–9], see Fig. 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. Ehrlich and Lacey found that studying hemimethylation provided valuable insight into cancer-associated active demethylation [5]. Exactly how the different hemimethylation patterns affect gene expression is not yet well documented, except for a recent finding that correlates gene body hemimethylation with an increase in transcription [10].

Hemimethylation may be closely related to hypermethylation and hypomethylation patterns found in the genome. When hypermethylation occurs in CpG islands proximal to the promoter regions of a gene, it is more likely to repress the gene [11]. In contrast, DNA hypomethylation is observed when CpG sites lose the methyl group and become less methylated or bare. Hypomethylation or unmethylation may be associated with increased gene expression [4].

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. This identification is useful because many patients are not diagnosed until the tumor is at an incurable stage [12]. In addition, DNA methylation patterns and levels can vary as cancer progresses [7]. For example, 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 such indicators, including hemimethylation, could allow for better comprehension of certain cancers and provide better insight toward the development of treatment methods. Furthermore, the possibility exists for future manipulation of methylation levels to interact with epigenetic changes associated with cancer and imprinting disorders [13].

Identifying the methylation pattern difference between normal and tumorous cells may indicate a given set of genes influencing carcinogenic development. Hemimethylation, the transitional state of DNA methylation, may be correlated with gene expression levels of certain genes. The hemimethylation of carcinogenic genes may be related to cancer development and progression. Therefore, it is important to investigate cancer hemimethylation. Hemimethylation has been researched previously for breast cancer cell lines [8], but not yet for lung cancer. Lung cancer is a great candidate for this investigation as it is challenging to detect. Its symptoms are often obscure or mistakable due to the consequence of previous smoking habits. Utilizing hemimethylation markers 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 lung cells using publicly available methylation sequencing data. In the Methods section, the statistical and computational analysis will be described, as well as the hemimethylation terminologies used throughout the paper. In the Results section, the outcomes of a variety of comparisons are evaluated and displayed, followed by biological mapping of highly hemimethylated genes and analysis of their relationships to cellular functions. All of these analyses guide towards the discussion of hemimethylation patterns in carcinogenic development and the conclusion of this study.

Methods

In this study, samples are obtained from lung 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 [14]. Sequencing reads are aligned to the hg38 reference genome and methylation signals are obtained using the BRAT-bw software package [15]. All methylation datasets consist of the methylation level of each CpG site, based on the number of sequencing reads that are methylated and unmethylated. Further analysis is then performed on CpG sites that exhibit at least 4 samples with non-NA methylation levels for both strands.

The Wilcoxon signed rank test is utilized to investigate if hemimethylation exists at each CpG site. 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 there is no methylation level difference between the forward and reverse strands at each CpG site.

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 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. For those in clusters, the tumor and normal strands 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. The CpG sites that are not in clusters are called singletons. There are similar numbers of hemimethylation sites in tumor and normal samples, but the proportion in clusters is slightly higher in normal samples. 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.

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.

|Mean difference|	Normal			Tumor		
	Total	Sites in clusters	Percentage	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%

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).

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.

	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

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.

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.

Cluster Pattern	Normal	Tumor
MMMMMMMMMMMM-UUUUUUUUUUUU	1	1
MMMMMMMMMM-UUUUUUUUUU	1	1
MMMMMMMM-UUUUUUUU	2	2
MMMMMMM-UUUUUUU	2	2
MMMMMM-UUUUUU	5	3
MMMMM-UUUUU	6	7
MMMM-UUUU	18	13
MMM-UUU	55	32
MM-UU	168	153
MMU-UUM	0	1
MU-UM	28	32
UMM-MUU	1	0
UM-MU	7	4
UUM-MMU	1	0
UU-MM	195	172
UUU-MMM	52	44
UUUU-MMMM	22	22
UUUUU-MMMMM	9	14
UUUUUU-MMMMMM	3	4
UUUUUUM-MMMMMMU	0	1
UUUUUUU-MMMMMMM	4	3
UUUUUUUM-MMMMMMMU	1	0
UUUUUUUU-MMMMMMMM	2	2
Total	583	513

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 Fig. 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.

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

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

Table 5
Polarity clusters with corresponding percentages.

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

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. 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 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.

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.

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

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.

Table 7
Hemimethylation frequency measured in gene bodies for both tumor and normal samples.

No. of Hemimethylation sites per gene body	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**>=10**
Tumor	1133	250	79	37	17	4	7	2	0	4
Normal	1118	229	73	32	11	4	3	1	1	5

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

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

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. In each of these tables, the first column is the gene name, the second column is the number of hemimethylation sites belonging to this gene, and the third column is the description of this gene. Terms in the tables that are followed by * are gene families, e.g., transcription factor and oncogene families. Otherwise they are general gene descriptions. The description and gene family of each gene are summarized based on the Molecular Signature Database [16] and the GeneCards database [17] .

There are 41 genes with the most hemimethylation in tumor DNA, see Table 9. Among these genes, TP73 [18–20], GNAS [21–25], and NOTCH1 [26, 27] are notable ones with known relations to cancer. Table 9 shows that among these 41 genes, 1 is a tumor suppressor (WT1), 3 are oncogenes (GNAS, NOTCh1, and PRDM16), and of those 3, 2 are translocated cancer genes (NOTCH1 and PRDM16). There are also 8 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, 4 are oncogenes (CBFA2T3, GNAS, PDGFB and PRDM16). Of the oncogenes, 3 are translocated cancer genes (CBFA2T3, PDGFB and PRDM16). There are also 7 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. Table 11 shows 36 genes with the most hemimethylation in both normal and tumor DNA. Among these genes, 2 are oncogenes and also translocated cancer genes (CBFA2T3 and PRDM16). There are also 6 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.

Table 9

Genes with ≥ 5 hemimethylation sites in tumor samples. The "*" beside certain genes indicates a specific gene family (e.g., transcription factor or oncogene family) that a gene belongs to.

Gene name	Count	Gene Description
RGS14	17	regulator of G protein signaling 14
MEX3A	16	mex-3 RNA binding family member A
WT1*	11	tumor suppressor, transcription factor*, WT1 transcription factor
PRDM16*	10	oncogene, translocated cancer gene, transcription factor*, 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	oncogene*, 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	transcription factor*, tumor protein p73
IFT140	6	intraflagellar transport 140
NFATC1*	6	transcription factor*, 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	transcription factor*, histone deacetylase 4
IRX2*	5	homeodomain protein, transcription factor*, iroquois homeobox 2

Gene name	Count	Gene Description
ANKRD33B	5	ankyrin repeat domain 33B
LINC00537	5	Long Intergenic Non-Protein Coding RNA 537
NOTCH1*	5	oncogene, translocated cancer gene*, notch receptor 1
ANO2	5	anoctamin 2
CACNA1H	5	calcium voltage-gated channel subunit alpha1 H
RUNX3*	5	transcription factor*, runt related transcription factor 3
SIX3*	5	homeodomain protein, transcription factor*, 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	protein kinase*, mitogen-activated protein kinase 1
SYN1	5	synapsin I

Table 10

Genes with ≥ 5 hemimethylation sites in normal samples. The "*" beside certain genes indicates a specific gene family (e.g., the transcription factor family) that a gene belongs to.

Gene name	Count	Gene Description
ZFPM1*	14	transcription factor*, zinc finger protein, FOG family member 1
GNAS*	13	oncogene*, GNAS complex locus
RGPD2	12	RANBP2 like and GRIP domain containing 2
SHANK3	11	SH3 and multiple ankyrin repeat domains 3
IRX2*	10	homeodomain protein, transcription factor*, 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	homeodomain protein, transcription factor*, homeobox A3
CACNA1H	6	calcium voltage-gated channel subunit alpha1 H
NPEPPS	6	aminopeptidase puromycin sensitive
SEMA6B*	6	cytokine or growth factor*, 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	transcription factor, oncogene, translocated cancer gene*, PR/SET domain 16
PLK3*	5	protein kinase*, polo like kinase 3
GREM2*	5	cytokine or growth factor*, gremlin 2, DAN family BMP antagonist
MEIS1*	5	homeodomain protein, transcription factor*, Meis homeobox 1

Gene name	Count	Gene Description
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	transcription factor, oncogene, translocated cancer gene*, CBFA2/RUNX1 translocation partner 3
RPL13	5	ribosomal protein L13
NFIC*	5	transcription factor*, nuclear factor I C
CDH4	5	cadherin 4
PDGFB*	5	cytokine or growth factor, oncogene, translocated cancer gene*, platelet derived growth factor subunit B
CCNT1	5	cyclin T1
SNORD68	5	small nucleolar RNA, C/D box 68

Table 11

Genes with ≥ 5 hemimethylation sites identical in both tumor and normal samples. The "*" beside certain genes indicates a specific gene family (e.g., transcription factor or oncogene family) that a gene belongs to.

Gene name	Count	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
SGTA	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
HOXA3*	6	homeodomain protein, transcription factor*, 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	oncogene, translocated cancer gene, transcription factor*, PR/SET domain 16
RGPD4	5	RANBP2 like and GRIP domain containing 4
MERTK*	5	protein kinase*, MER proto-oncogene, tyrosine kinase
FAM160A1	5	family with sequence similarity 160 member A1
PRKAR1B	5	protein kinase cAMP-dependent type I regulatory subunit beta
MAD1L1	5	mitotic arrest deficient 1 like 1
HOXA2*	5	homeodomain protein, transcription factor*, homeobox A2
DPP6	5	dipeptidyl peptidase like 6
DIP2C	5	disco interacting protein 2 homolog C

Gene name	Count	Gene Description
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	transcription factor*, Kruppel like factor 5
ISL2*	5	homeodomain protein, transcription factor*, ISL LIM homeobox 2
CBFA2T3*	5	oncogene, translocated cancer gene, transcription factor*, CBFA2/RUNX1 translocation partner 3
SBNO2	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

In order to understand the functions and relationships of these genes, we further analyze their biological interactions using the ConsensusPath Database (CPDB) software package [28–30], see Figs. 3, 4, 5, and 6. Figure 3 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 3 is the legend for Figs. 4, 5, and 6. This legend figure summarizes the relationships for gene lists in Tables 9, 10, and 11 as shown in Figs. 4, 5, and 6, respectively. These figures show the extent to which these highly hemimethylated genes interact and possibly affect the cell function of related genes.

Figure 4 shows genetic interactions between genes with the most hemimethylation in tumor samples, and these genes are recorded in Table 9. The gene network in Fig. 4 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 [31–34] and UNC5B [35–38], are known to be associated with cancer.

Figure 5 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 [8]. MEIS1 is also a

hub gene that interacts with genes like KMT2A [39] and TK1 [40]. While these genes are not hemimethylated in our samples, they are known to be associated with cancer. MEIS1 plays a crucial role in normal development [17] and it is also reported as an important gene related to leukemia [41–43]. 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 6 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 are 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 Fig. 4. There appears to be no common genes between Fig. 5 (hemimethylated genes in normal samples) and Fig. 6 (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.

Discussion

The original sequencing datasets are generated via the reduced representation bisulfite sequencing protocol. Because this sequencing method covers only a small percentage of the whole genome, there are many NA entries as shown in Table 2. A more thorough sequencing method like whole-genome bisulfite sequencing, which can provide methylation signals on all CpG sites in a genome, will help us see a clear picture of hemimethylation patterns in an entire genome.

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 [8]. 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 [44].

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. [8]. 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 [8]. 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 Fig. 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.

Conclusion

In this paper, we have conducted the first-ever in-depth investigation of hemimethylation for lung cancer. In particular, we have conducted statistical analyses to identify hemimethylation patterns for 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.

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.

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 is conducted at Texas State University with the support of NSF-REU grant DMS-1757233 for the three student authors. This project is supported by the Texas State University Research Enhancement Program (an internal award for Dr. Sun).

Authors' contributions

SS initiated the project, suggested all key original ideas, and led the whole process. All authors contributed to the early coding in the beginning of this project. AZ contributed to most of the statistical analysis. CF helped with the data analysis and JP did partial coding summary. 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.

References

1. **American Cancer Society (www.cancer.org)**. Assessed 27 January 2020.
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, Dubeau L, Ehrlich M: **Hemimethylation footprints of DNA demethylation in cancer**. *Epigenetics : official journal of the DNA Methylation Society* 2009, **4**(3):165-175.
8. Sun S, Lee YR, Enfield B: **Hemimethylation Patterns in Breast Cancer Cell Lines**. *Cancer Inform* 2019, **18**:1176935119872959.
9. Sun S, Li P: **HMPL: A Pipeline for Identifying Hemimethylation Patterns by Comparing Two Samples**. *Cancer Inform* 2015, **14**(Suppl 2):235-245.
10. Xu C, Corces VG: **Nascent DNA methylome mapping reveals inheritance of hemimethylation at CTCF/cohesin sites**. *Science* 2018, **359**(6380):1166-1170.
11. Esteller M: **CpG island hypermethylation and tumor suppressor genes: a booming present, a brighter future**. *Oncogene* 2002, **21**(35):5427-5440.
12. Walters S, Maringe C, Coleman MP, Peake MD, Butler J, Young N, Bergstrom S, Hanna L, Jakobsen E, Kolbeck K *et al*: **Lung cancer survival and stage at diagnosis in Australia, Canada, Denmark, Norway, Sweden and the UK: a population-based study, 2004-2007**. *Thorax* 2013, **68**(6):551-564.
13. Yang X, Yan L, Davidson NE: **DNA methylation in breast cancer**. *Endocr Relat Cancer* 2001, **8**(2):115-127.
14. 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.
15. 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.
16. 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 U S A* 2005, **102**(43):15545-15550.
17. **GeneCards - Gene Database (www.genecards.org)** Assessed 27 January 2020.
18. 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.

19. 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.
20. Ye H, Guo X: **TP73 is a credible biomarker for predicting clinical progression and prognosis in cervical cancer patients**. *Biosci Rep* 2019, **39**(8).
21. Hollstein PE, Shaw RJ: **GNAS shifts metabolism in pancreatic cancer**. *Nat Cell Biol* 2018, **20**(7):740-741.
22. 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.
23. 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.
24. 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.
25. Tominaga E, Tsuda H, Arao T, Nishimura S, Takano M, Kataoka F, Nomura H, Hirasawa 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.
26. 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.
27. 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.
28. 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.
29. Kamburov A, Stelzl U, Lehrach H, Herwig R: **The ConsensusPathDB interaction database: 2013 update**. *Nucleic acids research* 2013, **41**(Database issue):D793-800.
30. 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.
31. 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.
32. 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.
33. 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.

34. 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.
35. 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.
36. 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.
37. Liu J, Zhang Z, Li ZH, Kong CZ: **Clinical significance of UNC5B expression in bladder cancer**. *Tumour Biol* 2013, **34**(4):2099-2108.
38. 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.
39. Rao RC, Dou Y: **Hijacked in cancer: the KMT2 (MLL) family of methyltransferases**. *Nat Rev Cancer* 2015, **15**(6):334-346.
40. 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.
41. 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.
42. 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.
43. 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.
44. 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.

Figures

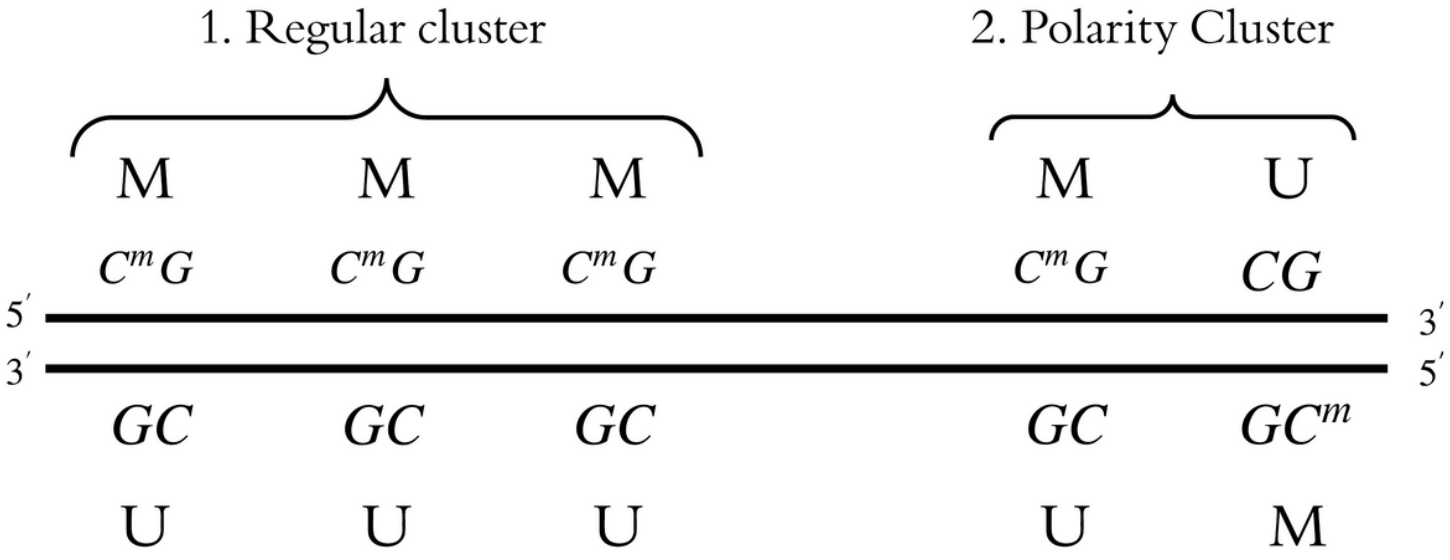

Figure 1

Examples of regular and polarity clusters shown on forward and reverse strands. CmG (or GCm) refers to a methylated (M) site; CG (or GC) refers to an unmethylated (U) site.

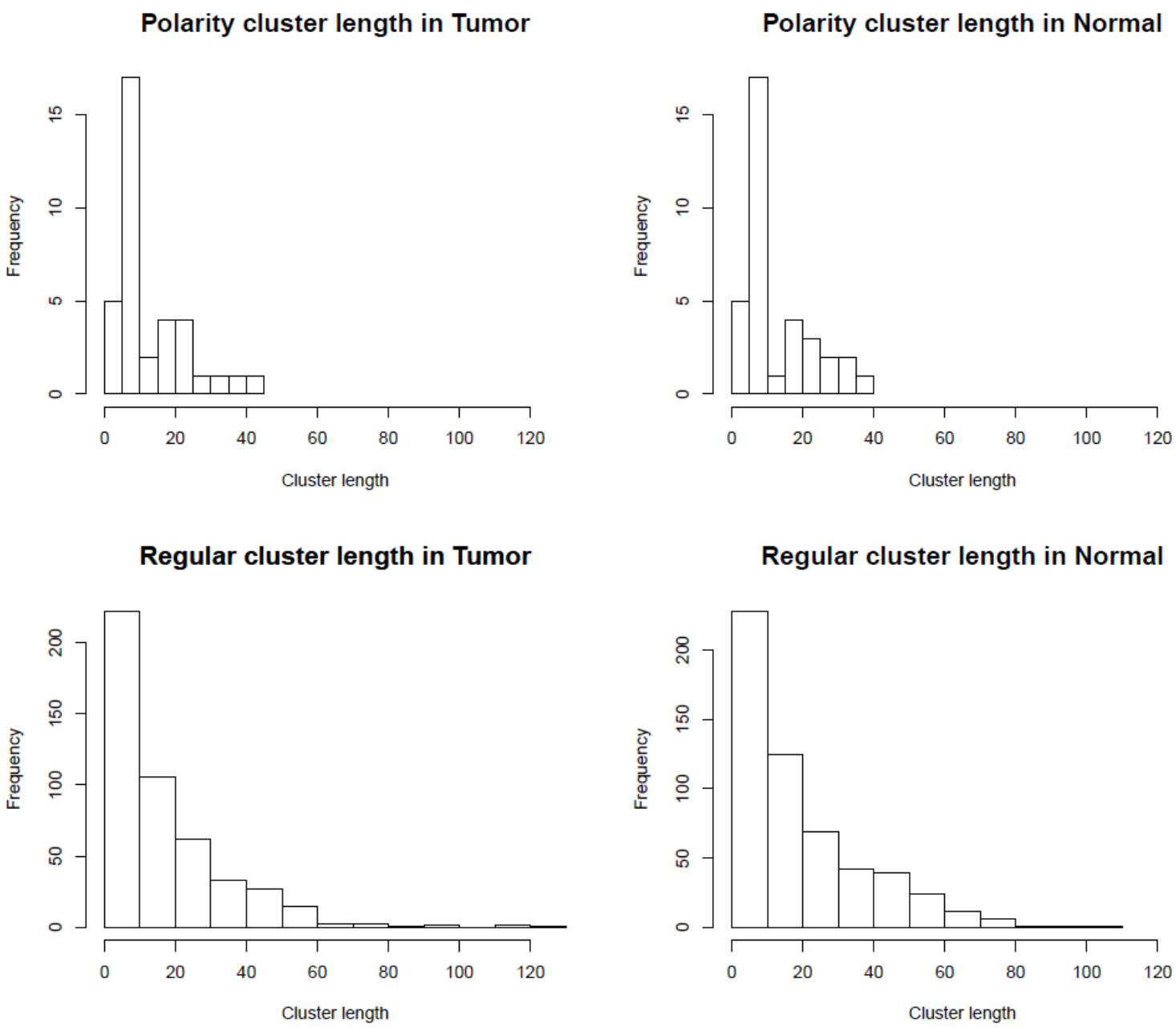

Figure 2

Length of clusters for both normal and tumor samples.

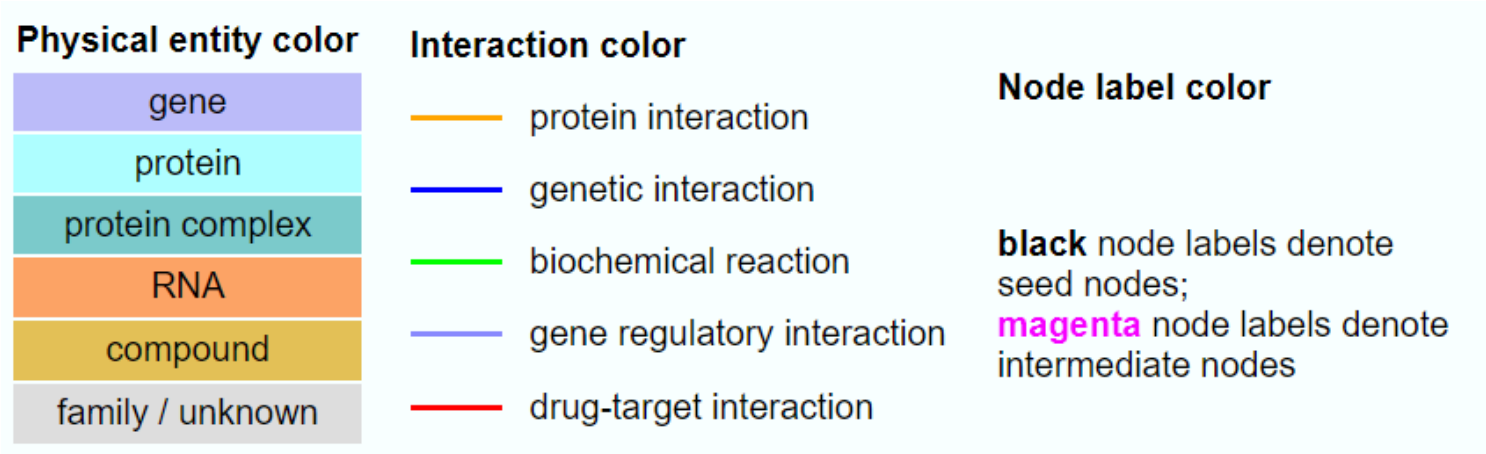

Figure 3

Key for gene relationship diagrams in Figures 4, 5, and 6.

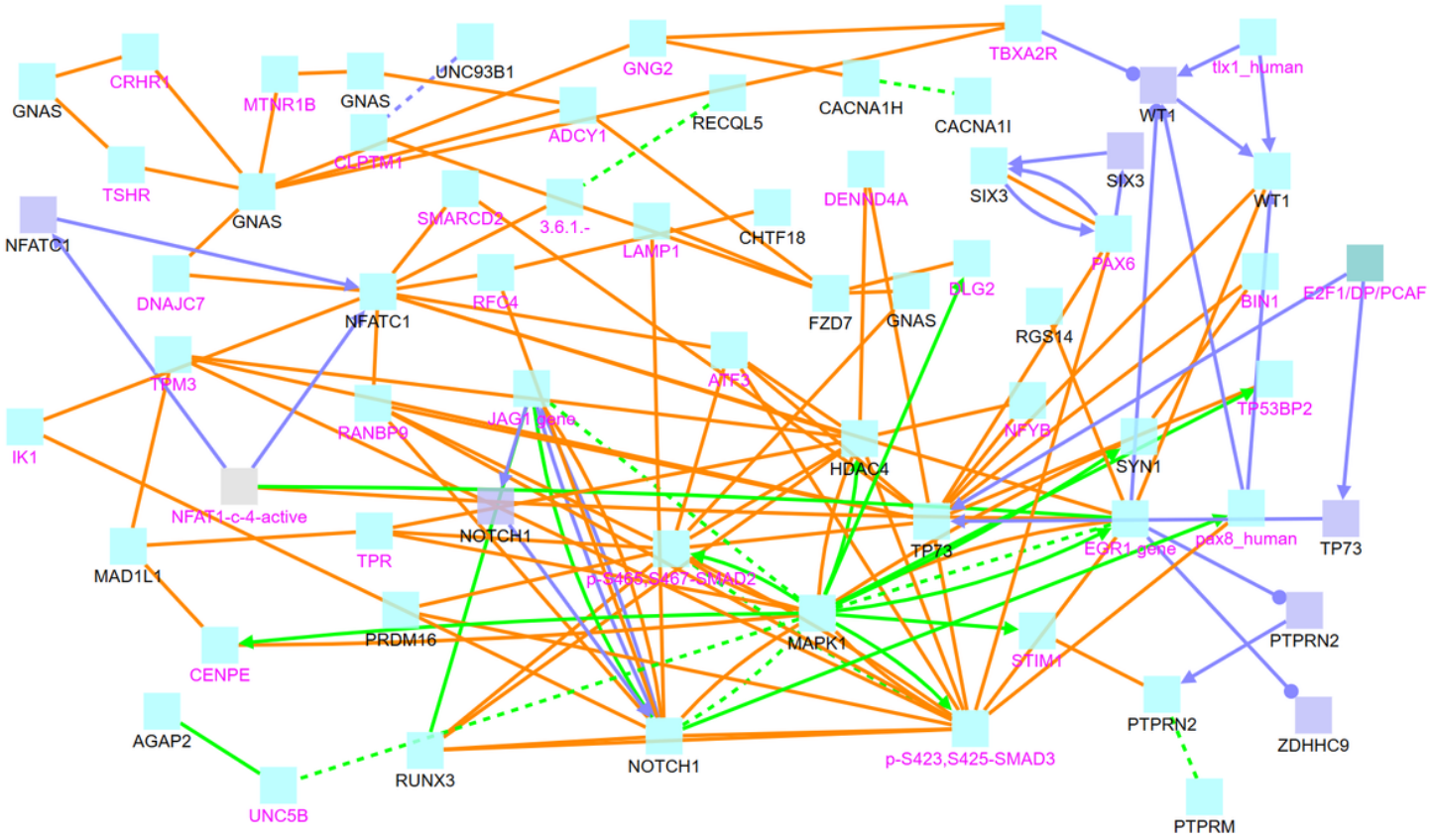

Figure 4

Relationship between genes with ≥ 5 hemimethylation sites in tumor samples.

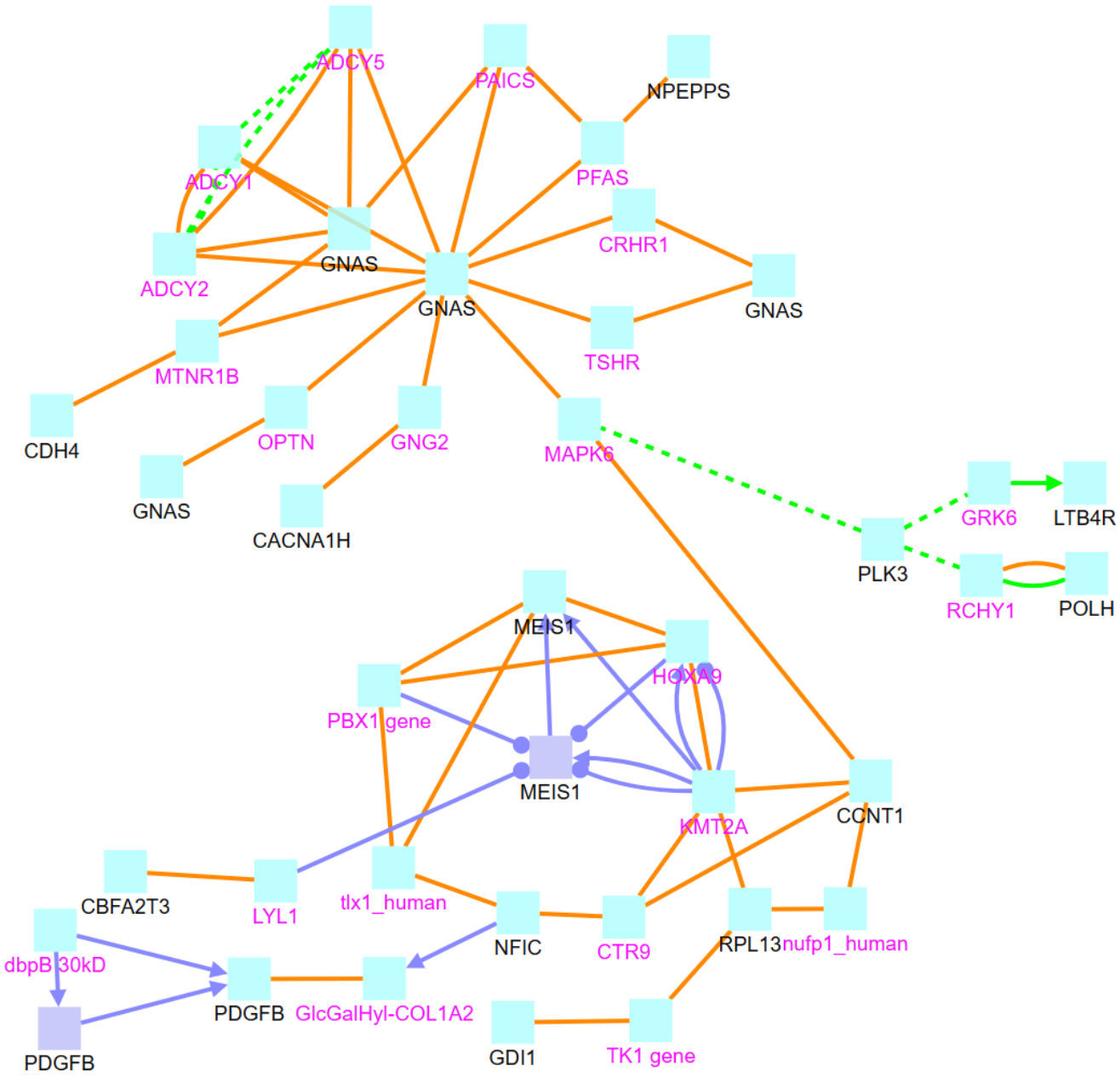

Figure 5

Relationship between genes with ≥ 5 Hemimethylation sites in normal samples.

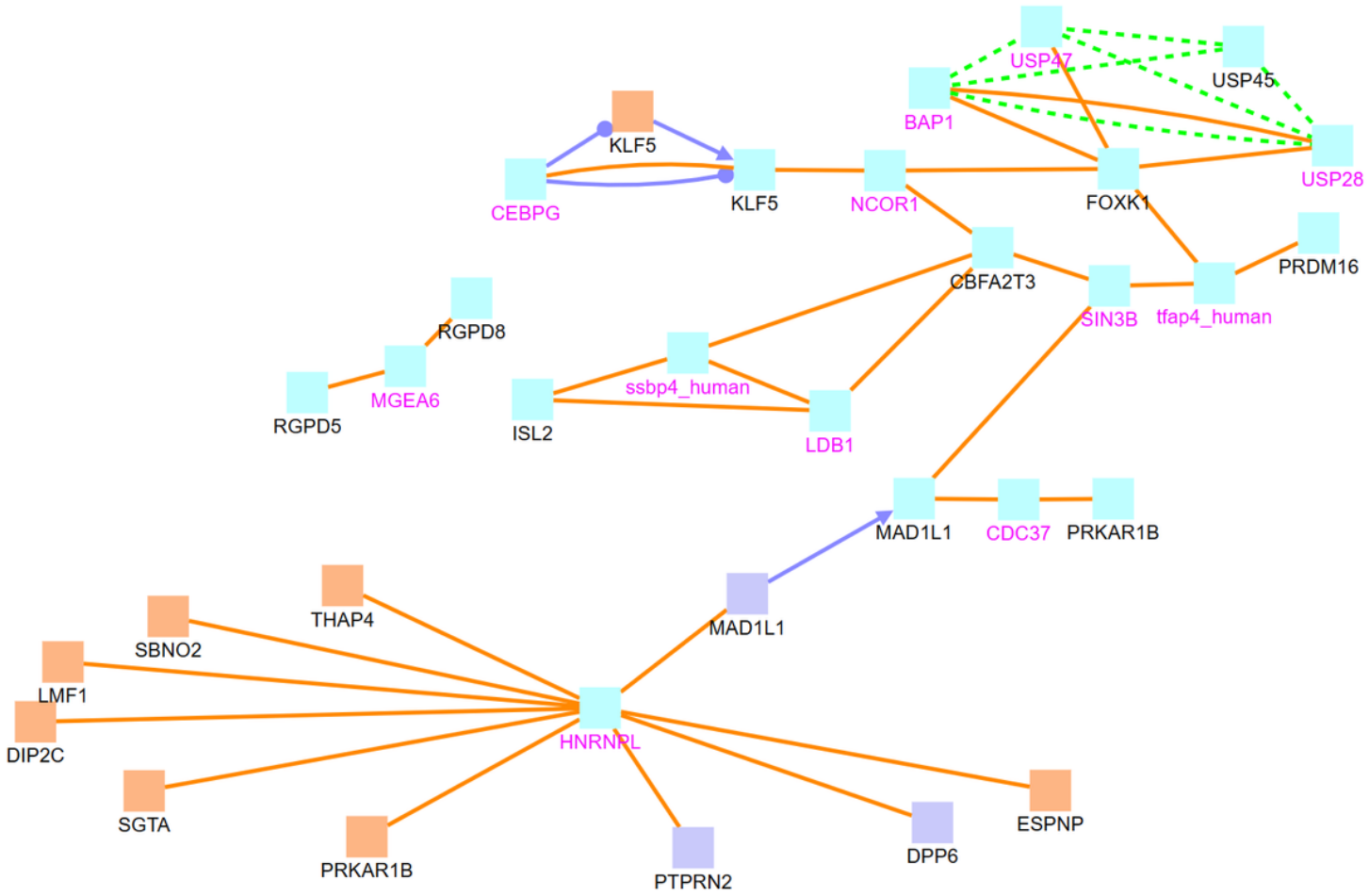

Figure 6

Relationship between genes with ≥ 5 hemimethylation sites identical in both tumor and normal samples.