



Correlation Patterns Between DNA Methylation and Gene Expression in The Cancer Genome Atlas

John CG Spainhour^{1,*}, Hong Seo Lim^{1,*} , Soojin V Yi² and Peng Qiu¹ 

¹Department of Biomedical Engineering, Georgia Institute of Technology and Emory University, Atlanta, GA, USA. ²School of Biological Sciences, Georgia Institute of Technology, Atlanta, GA, USA.

Cancer Informatics
Volume 18: 1–11
© The Author(s) 2019
Article reuse guidelines:
sagepub.com/journals-permissions
DOI: 10.1177/1176935119828776



ABSTRACT

BACKGROUND: DNA methylation is a form of epigenetic modification that has been shown to play a significant role in gene regulation. In cancer, DNA methylation plays an important role by regulating the expression of oncogenes. The role of DNA methylation in the onset and progression of various cancer types is now being elucidated as more large-scale data become available. The Cancer Genome Atlas (TCGA) provides a wealth of information for the analysis of various molecular aspects of cancer genetics. Gene expression data and DNA methylation data from TCGA have been used for a variety of studies. A traditional understanding of the effects of DNA methylation on gene expression has linked methylation of CpG sites in the gene promoter region with the decrease in gene expression. Recent studies have begun to expand this traditional role of DNA methylation.

RESULTS: Here we present a pan-cancer analysis of correlation patterns between CpG methylation and gene expression. Using matching patient data from TCGA, 33 cancer-specific correlations were calculated for each CpG site and the expression level of its corresponding gene. These correlations were used to identify patterns on a per-site basis as well as patterns of methylation across the gene body. Using these identified patterns, we found genes that contain conflicting methylation signals beyond the commonly accepted association between the promoter region methylation and silencing of gene expression. Beyond gene body methylation in whole, we examined individual CpG sites and show that, even in the same gene body, some sites can have a contradictory effect on gene expression in cancers.

CONCLUSIONS: We observed that within promoter regions there was a substantial amount of positive correlation between methylation and gene expression, which contradicts the commonly accepted association. We observed that the correlation between CpG methylation and gene expression does not exhibit in a tissue-specific manner, suggesting that the effects of methylation on gene expression are largely tissue independent. The analysis of correlation associated with the location of the CpG site in the gene body has led to the identification of several different methylation patterns that affect gene expression, and several examples of methylation activating gene expression were observed. Distinctly opposing or conflicting effects were seen in close proximity on the gene body, where negative and positive correlations were seen at the neighboring CpG sites.

KEYWORDS: CpG methylation, TCGA, pan-cancer, correlation, epigenetics

RECEIVED: January 3, 2019. **ACCEPTED:** January 9, 2019.

TYPE: Original Research

FUNDING: The author(s) disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: This study was partially supported by the National Science Foundation (CCF1552784) and the Giglio Family Breast Cancer Fund. P.Q. is an ISAC Marylou Ingram Scholar.

DECLARATION OF CONFLICTING INTERESTS: The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

CORRESPONDING AUTHOR: Peng Qiu, Department of Biomedical Engineering, Georgia Institute of Technology and Emory University, 950 Atlantic Dr NW, Atlanta, GA 30332-0230, USA.
Email: peng.qiu@bme.gatech.edu

Background

DNA methylation has been shown to play an important role in gene regulation and implicated in various types of cancer^{1–3} as well as the pathology of other medical conditions.⁴ The role of DNA methylation in cancer is particularly well appreciated, with numerous examples of cancer-specific CpG hypermethylation that turn off the expression of tumor suppressors and hypomethylation to activate the expression of oncogenes.^{5–9} As the extensive cancer databases such as the Cancer Genome Atlas (TCGA) have become readily accessible, our understanding of the role of DNA methylation can be confirmed and even expanded. TCGA provides comprehensive molecular characterizations of cancer samples, as well as clinical data of the corresponding patients, across 33 cancer types and a total

of ~11 000 patients. The molecular characterizations include mutation, copy number, methylation, gene expression, and protein expression.^{1,10–12} Such a data resource allows pan-cancer analysis which may reveal conserved and distinct patterns in different cancer types. For example, previous studies have used Pearson's correlation to examine the role of CpG methylation in gene expression. Pearson's correlation is a useful method for determining a possible link between different pieces of the gene regulatory network. Traditionally epigenetic modifications in cancer, such as the CpG island methylator phenotype (CIMP), have been shown to play a significant role in cancer pathology.^{12,13} However, thoughts on the role of methylation as solely a transcriptional silencer have begun to change to include the possibility of gene activation through changes in methylation.¹⁴ Current evidence points to the small subset of methylation sites that directly trigger gene expression.

*J.C.G.S. and H.S.L. are co-first authors.



Here we present a pan-cancer analysis to examine the association between DNA methylation and gene expression on an individual CpG basis. It is known that methylation of a single CpG can affect gene expression significantly.^{15,16} We examined individual CpGs along the entire gene body to evaluate CpGs' associations with gene expression using correlation analysis. We performed this analysis across 33 cancer types in the TCGA database. Due to this extensive analysis, we were able to further evaluate correlation patterns as they occur in different regions of the gene body. Specifically, the Transcription Start Site (TSS) and the Transcription End Site (TES) for each gene are used to categorize into 3 different regions relative to a gene. We divided each gene using TSS and TES so that the first region (region 1) is the upstream region before TSS, the second region (region 2) is in between TSS and TES, and the third region (region3) is the downstream region past the TES. This allows for the consideration of methylation sites in different genes to be readily comparable for all genes. In addition to dividing at TSS and TES, 2 additional locations, 2000bp upstream from TSS and 500bp downstream from TES, were used for the visualization purposes.

Furthermore, based on whether significant positive and negative correlations are observed in each of the 3 regions, a gene is assigned to 1 of 64 correlation patterns, which describes the overall effects of CpG methylation on the gene's expression. These 64 patterns are then placed in 3 general groups of patterns: consistent, long-range conflict, and short-range conflict.

We observed consistent patterns where the correlations are consistently positive or negative for all CpG sites associated with the gene. We also observed conflicting patterns where CpG sites in the same region exhibit opposite correlations to gene expression. Based on the observed prevalence of these correlation patterns, it appears that the methylation effects on gene expression are largely tissue independent, although methylation and expression themselves are often tissue specific. In addition, a small but significant portion of genes exhibit patterns indicating that they are regulated in a manner different from the traditional view of methylation silencing gene expression; There is a portion of these genes that show increased expression when methylated or show a conflicting effect where methylation sites close to one another have opposite effects on gene expression.

Methods

Data access and preprocessing

Both DNA methylation data and gene expression data from TCGA were accessed (2017 accession) through either the Genomic Data Commons (GDC) using the data portal¹⁷ or the data transfer tool TCGA-assembler 2.^{18,19} The TCGA-assembler downloads TCGA data in bulk, providing 1 file containing 1 data type (methylation or expression) for all patients in 1 cancer type. GDC downloads TCGA data in smaller pieces

and provides 1 file for 1 data type and 1 individual patient. When using GDC to access TCGA data, individual patient files were assembled into a per-cancer-type file using in-house scripts in an R computing environment.²⁰ Preprocessing consisted of patient and gene matching between data types, log transformation of gene expression data, and removal of all known single-nucleotide polymorphism (SNP)-associated CpG sites in the methylation data along with known gene fusion products. The methylation data in this study were acquired by the Illumina 450K array, which interrogates more than 450000 methylation sites on the Illumina chip. The data for this study contained information of 485578 CpG sites.

Instruction on acquiring and using the GDC data access tool can be found at the main GDC webpage (<https://gdc.cancer.gov/access-data/gdc-data-transfer-tool> accessed). TSS and TES information is downloaded through UCSC Genome browser. TSS and TES information of the reference genome of GRch37 (hg19) was used, and only the genes that have TSS and TES information are used for analysis.

Correlation analysis

Correlation analysis was performed using Pearson's correlation with a Bonferroni correction to the P -values based on the number of genes per cancer type shared between methylation and expression data sets. The correlation was performed between methylation beta values and log-base-2-transformed gene expression data with a Bonferroni-corrected p -value threshold of $\leq .05$. All statistical tests used standard R functions.

Define correlation patterns of methylation effects on expression

For each gene, $33 \times k$ correlations are calculated, where k is the number of CpG sites associated with the gene, and 33 is the number of cancer types available in TCGA. The TSS and TES for each gene are used to categorize into 3 separate regions relative to a gene. We divided each gene using TSS and TES so that the first region is the upstream region before TSS, the second region is in between TSS and TES, and the third region is the downstream region past the TES. This allows for the consideration of methylation sites in different genes to be readily comparable for all genes. In addition to dividing at TSS and TES, 2 additional locations, 2000bp upstream from TSS and 500bp downstream from TES, were used for the visualization purposes.

The exact locations of the x -axis (relative gene location) are evaluated as follows. Given a gene, if CpG site associated with the gene is located between TSS and TES, and given that the TSS and TES coordinates are known

$$Loc = \frac{CpG\ Coordinate - TSS}{TES - TSS}$$

Table 1. Top entropy scoring genes.

GENE	COUNT FOR EACH REGION	ENTROPY SCORE
SPATA17	[12 13 12 12 1 1]	2.998845536
SPATA1	[9 9 16 19 1 1]	2.994693795
SIDT1	[9 8 15 18 7 7]	2.991532758
KIF3B	[4 5 3 3 3 3]	2.99107606
RLTPR	[2 2 19 15 9 9]	2.989992792
FAM65A	[5 4 5 6 2 2]	2.985106271
FLRT2	[51 60 9 9 3 4]	2.980480685
ADAMTSL3	[11 10 7 5 4 4]	2.978232429
SETD1A	[13 11 6 8 28 25]	2.97790054
PXK	[3 3 30 21 1 1]	2.977417818
ANKAR	[7 5 7 8 1 1]	2.976660389
ATP2B4	[31 26 15 14 5 7]	2.973453185
AFF3	[25 19 85 68 20 16]	2.96869675
PDLIM2	[1 1 19 13 9 7]	2.963188812
FGFR3	[56 56 17 25 36 50]	2.954466429
RTEL1	[13 13 11 8 14 21]	2.952891381
CAMTA1	[18 14 19 14 320 223]	2.94893248
NOS1AP	[1 1 24 40 26 21]	2.9462548
MGAT5B	[2 2 40 24 49 39]	2.94509893
NTRK2	[3 2 6 7 2 3]	2.937628641

These genes show the most conflicting methylation signal across multiple cancer types. Of these only *FLRT2*, *RTEL1*, and *MGAT5B* have no previous known role in cancer.

If CpG site is located between TSS–2000bp and TSS

$$Loc = \frac{0.4 \times (CpG\ Coordinate - TSS)}{2000}$$

If CpG site is located prior to TSS–2000bp

$$Loc = -0.4 + \frac{-0.6 \times (TSS - CpG\ Coordinate - 2000)}{\text{maximum}(TSS - CpG\ Coordinate) - 2000}$$

If CpG site is located between TES and TES + 500bp

$$Loc = 1 + \frac{0.1 \times (CpG\ Coordinate - TES)}{500}$$

If CpG site is located after TES + 500bp

$$Loc = 1.1 + \frac{0.9 \times (CpG\ Coordinate - TES - 500)}{\text{maximum}(CpG\ Coordinate - TES) - 500}$$

We defined region 1 to be the region prior to TSS. Region 2 is the region between TSS and TES, and region 3 represents the region past the TES. For each of the 3 regions, the total numbers of significant positive and negative correlations are counted separately, as shown in Figure 2. The presence or absence of significant positive and negative correlations in the 3 regions can be encoded as a 6-digit binary identifier for a gene. In this binary identifier, a “1” signifies that there are significant correlations for a given gene between its expression and its CpG methylation in the given region. The first 3 digits denote positive correlations in each region and the last 3 denote negative correlations in each region. Such a binary identifier for each gene allows us to classify genes regarding their patterns of methylation-expression correlation.

There are 64 patterns that can be pulled from this construction to denote significant correlations in a given region of the gene length-correlation plot, as shown in Figure 2. In this study, only 63 of these patterns are considered because 1 pattern is the “empty pattern,” that is, [0 0 0 0 0], where the gene in question has no statistically significant correlations with any of its associated CpG sites. Correlation counts (Supplemental Table 1) are done for each gene for each cancer type separately,

and each gene is further categorized into small, medium, and large sizes depending on its gene length. TES–TSS is calculated for each and every gene, and then the 33rd and 66th percentiles were used to categorize into 3 different sizes.

Entropy analysis

Once the correlation counts, as shown in Figure 2(B), are obtained for all the genes, an entropy calculation was performed for each gene to determine the consistency of directions of the significant correlations. For each of the 3 regions, the counts for significant positive and negative correlations are normalized into a probability distribution, and an entropy score is calculated using the following formula

$$H = -P \log_2(P) - (1-P) \log_2(1-P)$$

where P is the percentage of significant positive correlations for a given region and $1-P$ is the percentage of significant negative correlations for that region. The entropy scores for the 3 regions are then summed to give an overall entropy score for a given gene. This overall entropy score is used to rank genes according to the consistency of directions of the significant correlations in the same gene region. This score has a range of 0 to 3, 0 for perfectly consistent cases, the significant correlations of which in each region are always of the same direction, and 3 for genes with an equal distribution of positive and negative correlations across all the 3 regions. A gene that has a low entropy score is one that methylation of its CpG sites affects gene expression similarly, whereas a gene that has a high entropy score shows a great variability in the effects of CpG methylation on gene expression.

Results

Correlations beyond promoter–methylation silencing expression

The correlation analysis in 33 cancer types led to 923 898 significant methylation–expression correlations associated to 17 415 genes after a Bonferroni correction for P -values $< .05$. Due to the variations in gene length and number of CpG sites associated to each gene, the number of significant correlations for each gene ranged from 1 per gene to a maximum of 4958 for *PRDM16*. This is after filtering out correlations associated with fusion gene products and any CpG sites that contain known SNPs. Fusion gene products are CpG sites that come from the fusion of 2 genes through translocation, interstitial deletion, or chromosomal inversion. These are removed to clean up data about the parent genes that are expressed individually and may be worth further examination in future work. When the significant methylation–expression correlation for all the genes is overlaid in Figure 1, we observed a spread of methylation effects across the entire gene body. In region 1, 70.5% of the correlations were negative. In region 3, however,

only 33.7% were negative. Thus, although the well-recognized negative correlation between promoter region methylation and gene expression was confirmed, yet non-trivial percentage, 29.5%, of positive correlations were discovered in the promoter region. Also, region 3, which represents the downstream past the TES, shows a greater number of positive correlations than negative correlations.

Methylation–expression correlation is tissue independent

Examining the methylation–expression correlations across multiple cancer types for a single gene illustrates different aspects of the effects of methylation on gene expression. One aspect is that, for a particular CpG site, its correlations with the corresponding gene expression were typically in the same direction for all the cancer types that exhibited significant correlations. Among the 220 641 CpG sites that have significant correlations, 61 438 (27.85%) always had positive correlation values, 118 474 (53.69%) always had negative correlation values, whereas 40 729 (18.46%) exhibited both positive and negative correlation values for all cancer types. One example is shown in Figure 2, where all methylations to gene expression correlations for *ABHD8* gene are plotted. Each vertical stripe corresponds with the correlations between 1 CpG site and its gene expression in 33 cancer types. The black circles indicate the significant correlations, and the gray circles indicate the insignificant ones. This observation that the black circles tend to be in the same direction is an indication that the correlation between methylation and gene expression is largely tissue independent, although studies have shown that methylation and expression themselves are tissue-dependent.²¹

The hierarchical clustering dendrograms in Figure 3 further support the general observation that the correlation between methylation and gene expression is tissue independent, whereas methylation and expression are each tissue dependent. The dendrogram generated from the average gene expression profile of patients in each cancer type clusters many well-known similar cancer types. LUAD (lung adenocarcinoma) and LUSC (lung squamous cell carcinoma), which are lung cancers, are clustered together and, KICH (kidney chromophobe), KIRC (kidney renal clear cell carcinoma), and KIRP (kidney renal papillary cell carcinoma), which are kidney cancers, and UCES (uterine corpus endometrial carcinoma) and UCS (uterine carcinosarcoma), which are uterus cancers, and GBM (glioblastoma multiforme) and LGG (brain lower grade glioma), which are brain cancers, are, respectively, clustered together. Overall, the dendrogram for gene expression shows strong tissue dependence. The dendrogram generated from the average methylation beta values of patients in each cancer type clusters some similar cancer types such as the brain, kidney, and uterus cancers, but not lung cancers. The tissue dependency is somewhat weaker in the

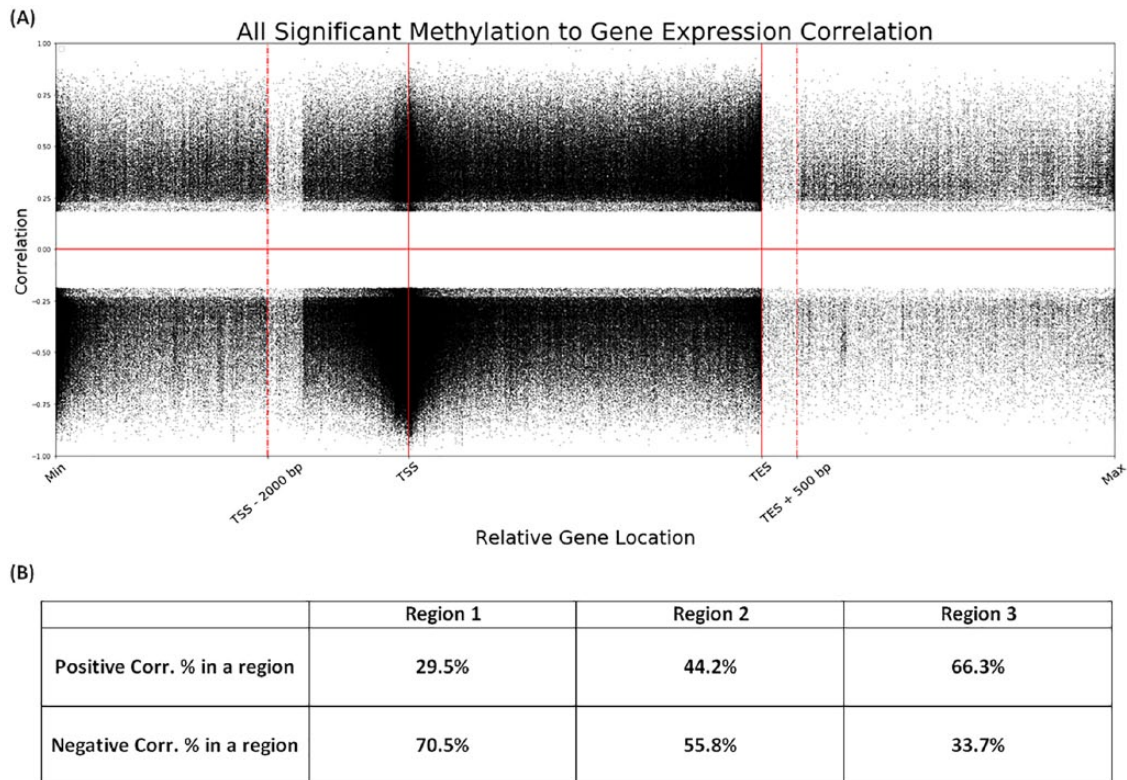


Figure 1. Total methylation to gene expression correlation results. (A) Total significant correlation results across all genes and cancer groups found in the study. The x-axis (relative location of CpG sites with respect to their associated gene) is divided into 3 regions based on TSS and TES of genes. Locations of TSS–2000bp and TES + 500 bp are represented as dotted lines for visual purposes. (B) A trend can be seen where correlations at region 3, past TES, tend to have more positive correlation (66.3%) with gene expression, whereas gene expression correlates negatively (70.5%) near the TSS. This shows evidence that the location of the CpG site influences the effect of methylation on transcription. TES: Transcription End Site; TSS: Transcription Start Site.

methylation data, yet the overall dendrogram analysis suggests that the methylation data are still tissue dependent. The dendrogram generated by methylation-expression correlations in each cancer at each CpG site, however, does not show patterns that were found in other 2 dendrograms. The only similar cancer types, which are aforementioned above, clustered together are brain cancers (GBM and LGG). Such observation once again suggests that the correlation between methylation and gene expression is largely tissue independent, contrary to methylation and expression, respectively, being tissue dependent. For example, the relationship between methylation and expression could follow a pattern defined early during development.^{22,23}

Nearby CpG sites can exhibit the opposite effect on expression

Another aspect is that CpG sites near one another often share the same sign regarding their correlation to gene expression in most of the cases. For example, in Figure 4, methylation of CpG sites near the TSS of NYNRIN gene has almost all negative correlations with the NYNRIN expression for all cancer types that showed significant correlations. However, there were also a non-trivial number of cases where CpG sites nearby each other can exhibit conflicting correlations with opposite signs. For example, in Figure 5, we observed that the

methylation-expression correlation of CpG sites in OSR1 jumps up and down even though the locations of the sites are nearby. For such drastically different correlations in nearby CpG sites, we call them short-range conflicts.

We also observed patterns we considered as long-range conflicts. As shown in Figure 6, methylation of CpG sites right before the TSS of ZNF282 negatively correlates with ZNF282 expression, whereas most methylation of CpG sites after the TES that are significant positively correlate with expression. Such a difference indicates the possibility that the methylation of CpG sites at different locations in the body of a gene have different regulatory roles or functions.

A variety of methylation-expression correlation patterns at gene level

When examining the methylation-expression correlation of multiple cancer types at the gene level, we observed a variety of correlation patterns, which necessitated visualizations to efficiently describe and summarize these patterns. One approach is to summarize the correlation effects of methylation over regions of the gene. Previous work^{21,24} has shown that methylation in different regions of the gene body affects gene expression differently. For our studies, each gene is divided into 3 regions as explained in the background section.

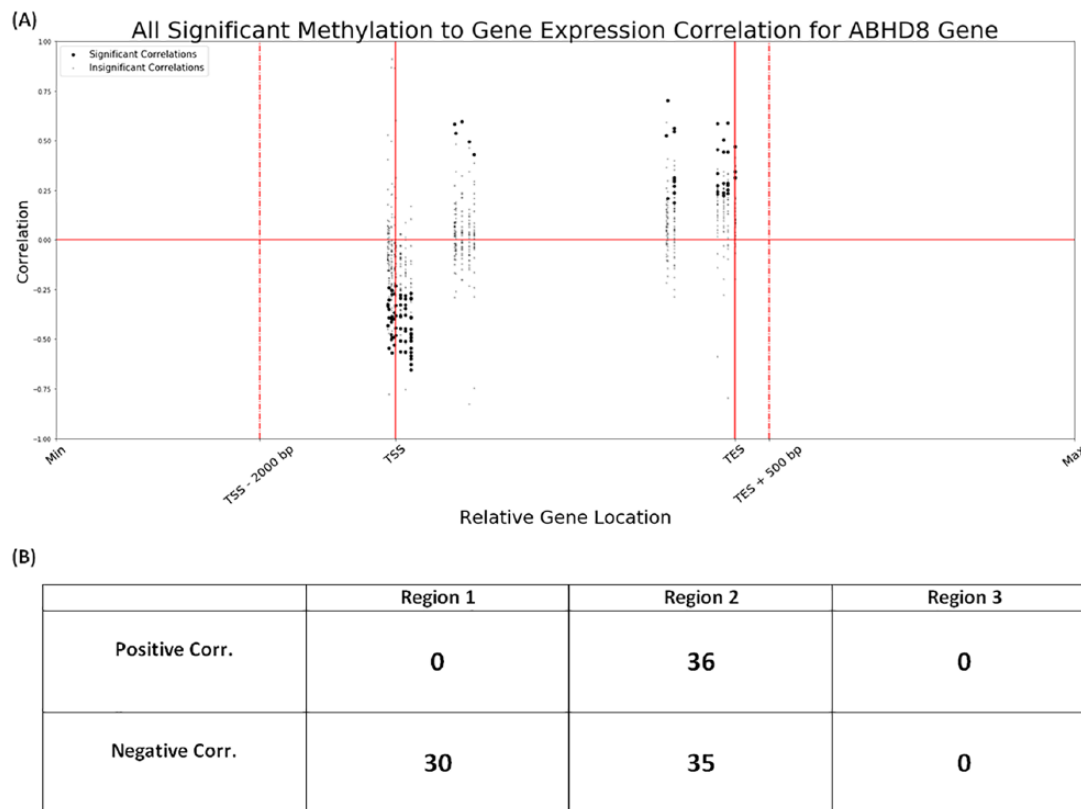


Figure 2. Plot of correlations to relative gene location and correlation pattern for ABHD8 gene across all cancers. (A) Statistically significant methylation-gene expression correlations are colored in thick black and all insignificant correlations are in gray. Each CpG site has 33 separate correlations, 1 for each cancer type in TCGA, which can be used to determine the overall methylation effect at the given CpG methylation site for that cancer. (B) The table represents the count of significant correlations found in the defined regions of the gene, separating positive from negative correlations. This serves as an example of short-range conflict where conflicting signals from methylation on gene transcription are observed in region 2. TCGA: The Cancer Genome Atlas.

After dividing a gene, for each cancer, we counted the number of significant positive and negative correlations for each region separately. The presence or absence of significant correlations in the 6 entries of this table forms a 6-digit binary code for a gene. There are 63 non-trivial correlation patterns in total, which can be organized into 3 categories: consistent, short-range conflicting, and long-range conflicting, as shown in Figure 7. Consistent patterns consist of genes that have correlations that are only positive or negative in one or more of the regions. Genes with long-range conflicting patterns have both positive and negative correlations but the positive and negative correlations are in separate regions. Short-range conflicting patterns refer to genes with both significant positive and negative correlations in the same region. These short-range conflicting genes appear to be more interesting, because the conflicting methylation signals indicated that the nearby CpG sites can have an opposite effect on gene expression.

For each gene, we examined its methylation-expression correlations in each of the 33 cancer types separately. One gene may exhibit different correlation patterns in the different cancer type. Figure 7 provides a summary of all the observed correlation pattern genes, organized into the 3 categories. The most prevalent patterns are the consistent ones, accounting for

73.8% of the cases. Among these, 53.6% cases show consistent negative correlations between methylation and expression, which fits the well-accepted mechanism that methylation silences transcription. In the meantime, 20.2% of cases showed consistently positive correlation (Figure 8(A)), illustrating that not all genes are affected by methylation in the same way. In addition, 5.0% showed long-range conflict (Figure 8(B)), where methylations across the gene body have a different effect on expression. Finally, 21.2% showed short-range conflict (Figure 8(C)) which may be of special interest, because they represent genes for which the methylation status changes drastically around nearby CpG sites, but changes in a way that strongly correlates with gene expression. Notice that the number of genes for each pattern is a cumulative number for all 33 cancers. As expected, the negative consistent pattern was most prevalent with 53.6%, but the fact that there are 46.4% of cases that show patterns that are contrary to the methylation causing silence of a gene expression suggests further studies are needed to grasp the complex mechanisms of interactions of methylation and gene expression. Also as shown in Figure 7, although it seems that short-range conflicting (21.2%) is more prevalent than long-range conflict (5.0%), because there are 37 different possible patterns for the short-range conflicts, compared with

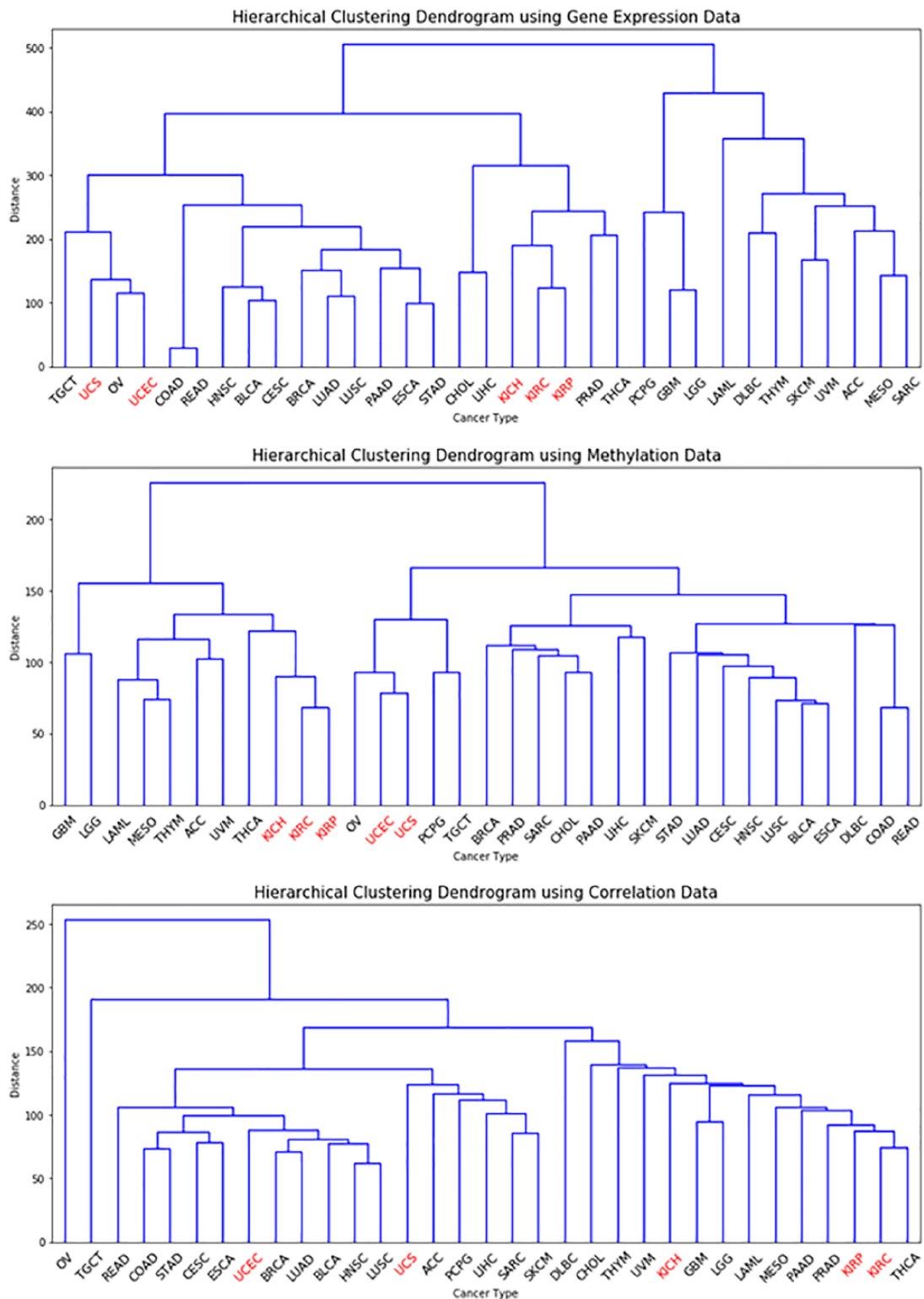


Figure 3. Hierarchical clustering dendrograms using different data sets. The Ward metric was used to perform hierarchical/agglomerative clustering on 3 datasets: gene expression, methylation, and correlation data. The dendrograms generated from gene expression (top panel) and methylation (middle panel) were able to cluster similar cancer, such as KICH, KIRC, and KIRP, which are all kidney cancers, and UCEC and UCS, which are uterus cancers. However, the dendrogram generated from methylation-to-gene expression correlation data (bottom panel) was not able to cluster the similar cancer as precisely as the other two.

KICH: kidney chromophobe; KIRC: kidney renal clear cell carcinoma; KIRP: kidney renal papillary cell carcinoma; UCEC: uterine corpus endometrial carcinoma; UCS: uterine carcinosarcoma.

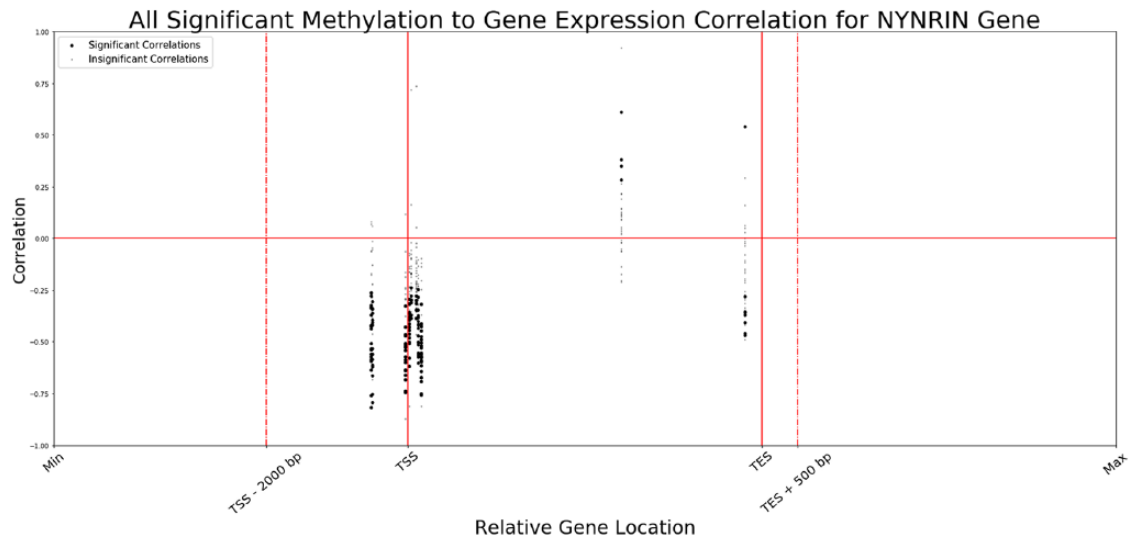


Figure 4. Plot of correlations to relative gene location and correlation pattern for NYNRIN gene across all cancers. This figure shows an example of the correlation plot where the CpG sites near TSS show consistent negative correlations. Gray circles denote all correlations and the black circles are significant correlations. TSS: Transcription Start Site.

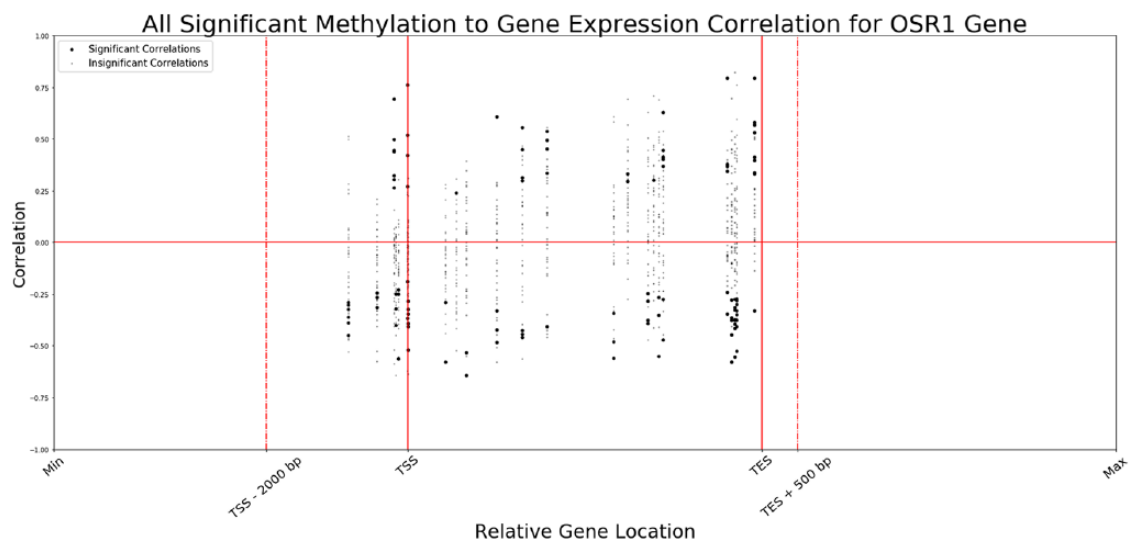


Figure 5. Plot of correlations to relative gene location and correlation pattern for OSR1 gene across all cancers. OSR1 shows multiple short- and long-range correlation conflicts. Contrary to NYNRIN gene, OSR1 gene shows alternating positive and negative correlations in even nearby CpG sites.

12 for the long-range conflicts, the short-range conflicts are relatively rare among the 4 patterns, which drew our attention.

Short-range conflicts are enriched by genes involved in cancers

To further examine the genes that exhibited short-range conflicting patterns of methylation-expression correlation, we used an entropy measure to rank the genes. Those genes with the most short-range conflict between methylation signals receive the highest entropy, calculated by the sum of the entropy of each individual region of a gene.

In Table 1, the 20 genes with the highest entropy are shown. These genes cover processes including cell motility, proliferation, and transcription. Out of the top 20 highest entropy

genes, 17 showed associations with cancer. For example, KIF3B, one of the top genes in Table 1, has been known to play an important role in hepatocellular carcinoma,²⁵ and FGFR3 is known to be associated with bladder cancer.²⁶ Among the top 20 genes in Table 1, only the 3 in bold (*FLRT2*, *RTEL1*, and *MGAT5B*) have no previous evidence of involvement in cancer. Because 17 of the top 20 highest entropy genes showed associations with cancer, it is our assumption that *FLRT2*, *RTEL1*, and *MGAT5B* could also be possibly related to cancer. *FLRT2* encodes molecules that regulate embryonic vascular development, *RTEL1* encodes a DNA helicase that manages telomeres, and *MGAT5B* encodes a protein for adhesion and migration of cells. The above genes' respective functionalities are often associated with cancer. It is, therefore, possible that

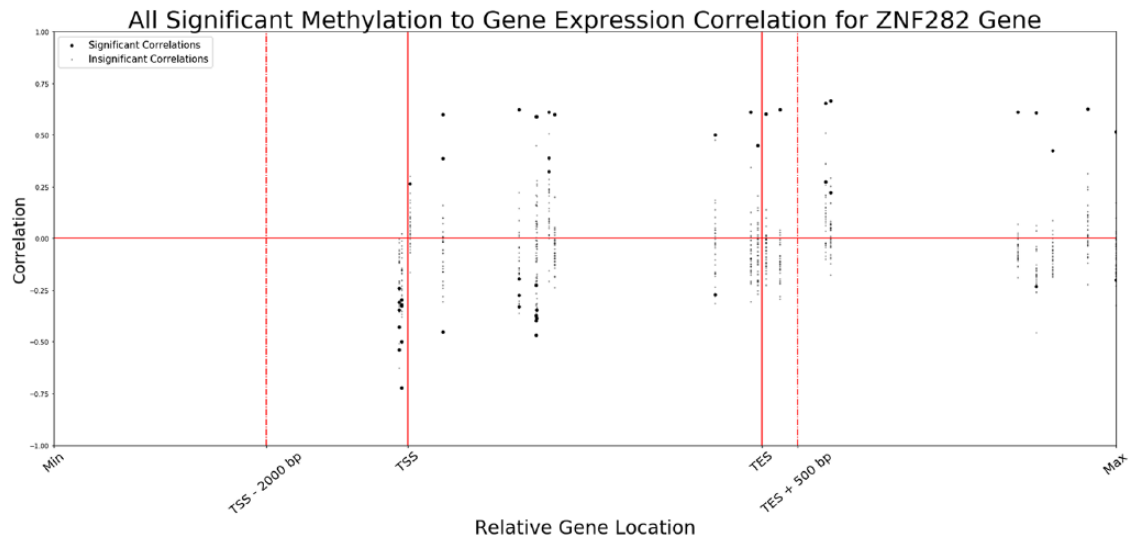


Figure 6. Plot of correlations to relative gene location and correlation pattern for ZNF282 gene across all cancers. ZNF282 shows a long-range correlation conflict. There exist significant negative correlations right before TSS (region 1), and there also exist positive correlations after TES (region 3). The difference in signs of correlations that are shown in different regions makes it a long-range conflict. TES: Transcription End Site; TSS: Transcription Start Site.

Pattern Type	Patterns	Number of Patterns	Number of Genes	Percentage
Positive Non-conflicting		7	44,421	20.2%
Negative Non-conflicting		7	117,929	53.6%
Long range Conflicting		12	10,963	5.0%
Short Range Conflicting		37	46,733	21.2%

Figure 7. Binary pattern used to examine the methylation activity. This table shows the types of methylation patterns found when using the grid system discussed. Non-conflicting patterns describe genes that have all positive or negative correlations in any combination of regions. Long-range conflicts refer to genes where positive and negative correlations are found in different regions. Short-range conflict describes genes that have positive and negative correlations in the same region, either on the same CpG site or on closely related CpG sites. The number of genes comes from correlations across multiple cancers found in TCGA and a gene is counted multiple times when it has correlations in different cancers. Note that most of the genes follow the traditional theory of gene methylation, where methylation is linked to silencing gene transcription. TCGA: The Cancer Genome Atlas.

these genes play a role in cancer and can serve as possible treatment targets after the future investigation.

Conclusions

This work represents an integrative pan-cancer analysis using TCGA data. By examining the correlation between methylation and gene expression, for various CpG sites and their corresponding genes, in various cancer types, we observed several different patterns of methylation-expression correlation. Whereas most of the genes display the expected correlation consistent with methylation-induced expression silencing, there is a significant proportion of genes that display patterns consistent with methylation-induced transcriptional

increases or a mixture dependent on the location of the methylation. Genes that showed significant conflicting effects were identified. The analysis across all 33 cancer types also shows that the effects of methylation on gene expression are largely tissue independent. These results clearly show that there is a great deal of more to be learned regarding the role of DNA methylation beyond the traditional silencing role. The methylation data in this study were acquired by the Illumina 450K array, which only interrogates CpG sites on the Illumina chip. One future direction could be to use sequencing-based methylation data to examine the methylation-expression correlation for CpGs that were not included in the 450K array.

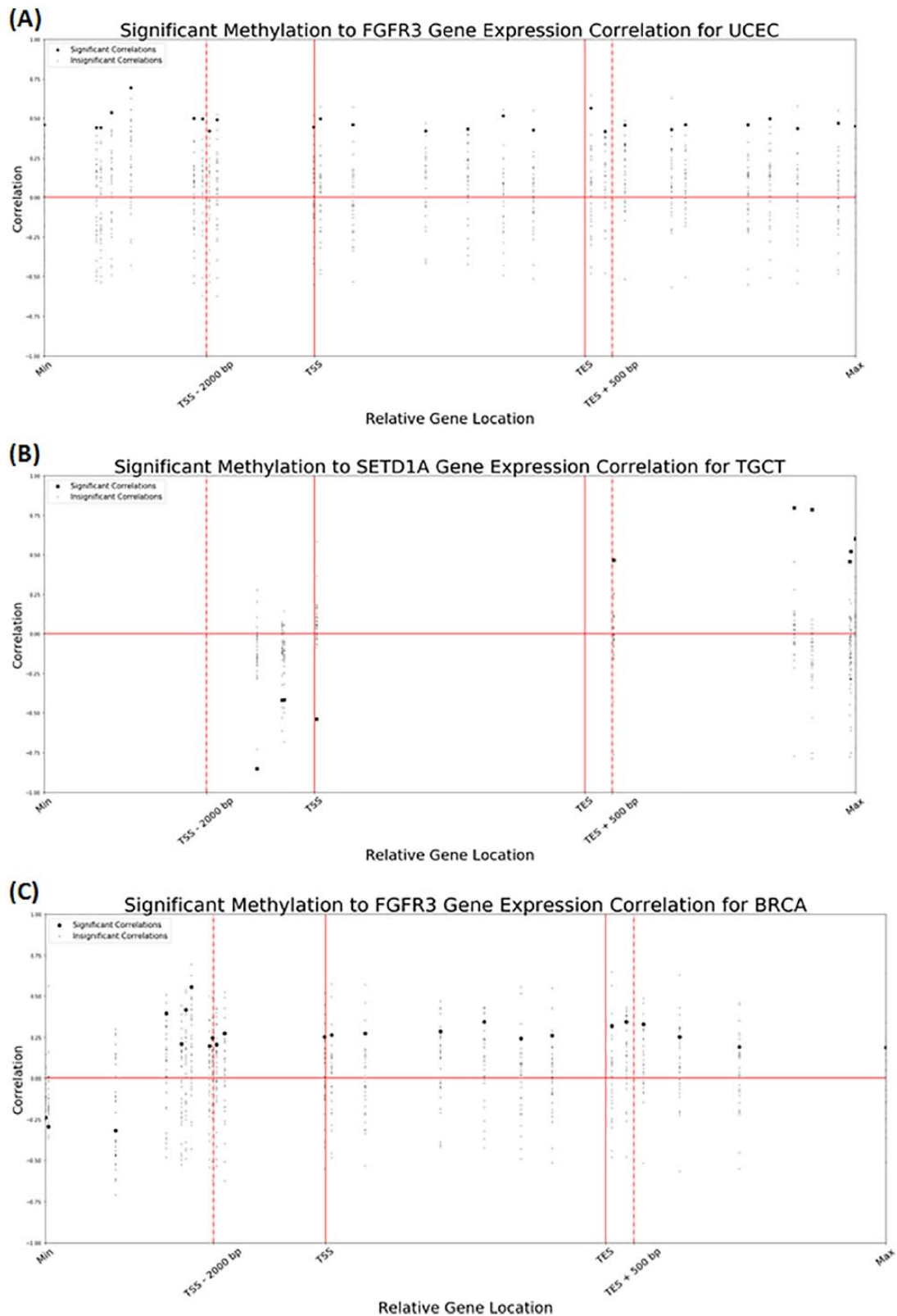


Figure 8. Methylation-expression correlations for specific cancers. (A) For UCEC (uterine corpus endometrial carcinoma), significant correlations for FGFR3 gene are plotted. Region 1, 2, and 3 all show positive correlations, so this type of pattern will belong to “positive non-conflicting.” (B) For TGCT (testicular germ cell cancer), significant correlations for SETD1A gene are plotted. Regions 1 and 2 show negative correlations, whereas region 3 shows positive correlations. This type of conflicting pattern will belong to “long-range conflicting.” (C) Significant correlations for BRCA (breast invasive carcinoma) are plotted. There exist negative correlations in region 1, whereas positive correlations are also found in region 1. This pattern will belong to “short-range conflicting.”

Author Contributions

JCGS and HSL performed the analysis and prepared the manuscript. SVY and PQ designed and supervised the project and reviewed the manuscript. All authors read and approved the final version of the manuscript.

Availability of Data and Material

All data used in this analysis can be found at the GDC data portal. Samples of code used in this analysis are included in the additional files.

ORCID iDs

Hong Seo Lim  <https://orcid.org/0000-0002-3641-231X>

Peng Qiu  <https://orcid.org/0000-0003-3256-0734>

Supplemental Material

Supplemental material for this article is available online.

REFERENCES

- Noushmehr H, Weisenberger DJ, Diefes K, et al. Identification of a CpG island methylator phenotype that defines a distinct subgroup of glioma. *Cancer Cell*. 2010;17:510–522.
- Kim M, Costello J. DNA methylation: an epigenetic mark of cellular memory. *Exp Mol Med*. 2017;49:e322.
- Su J, Huang Y, Cui X, et al. Homeobox oncogene activation by pan-cancer DNA hypermethylation. *Genome Biol*. 2018;19:108.
- Chen C, Zhang C, Cheng L, et al. Correlation between DNA methylation and gene expression in the brains of patients with bipolar disorder and schizophrenia. *Bipolar Disord*. 2014;16:790–799.
- Jones PA, Baylin SB. The epigenomics of cancer. *Cell*. 2007;128:683–692.
- Sproul D, Kitchen RR, Nestor CE, et al. Tissue of origin determines cancer-associated CpG island promoter hypermethylation patterns. *Genome Biol*. 2012;13:R84.
- Jones PA. Functions of DNA methylation: islands, start sites, gene bodies and beyond. *Nat Rev Genet*. 2012;13:484–492.
- Schübeler D. Function and information content of DNA methylation. *Nature*. 2015;517:321–326.
- The Cancer Genome Atlas Research Network. Comprehensive genomic characterization defines human glioblastoma genes and core pathways. *Nature*. 2013;494:506.
- Hoadley KA, Yau C, Wolf DM, et al. Multi-platform analysis of 12 cancer types reveals molecular classification within and across tissues-of-origin. *Cell*. 2014;158:929–944.
- Miller BF, Sánchez-Vega F, Elnitski L. The emergence of pan-cancer CIMP and its elusive interpretation. *Biomolecules*. 2016;6:E45.
- Toyota M, Ahuja N, Ohe-Toyota M, Herman JG, Baylin SB, Issa JP. CpG island methylator phenotype in colorectal cancer. *Proc Natl Acad Sci U S A*. 1999;96:8681–8686.
- Halpern KB, Vana T, Walker MD. Paradoxical role of DNA methylation in activation of FoxA2 gene expression during endoderm development. *J Biol Chem*. 2014;289:23882–23892.
- Nile CJ, Read RC, Akil M, Duff GW, Wilson AG. Methylation status of a single CpG site in the IL6 promoter is related to IL6 messenger RNA levels and rheumatoid arthritis. *Arthr Rheum*. 2008;58:2686–2693.
- Long M, Smiraglia D, Campbell M. The genomic impact of DNA CpG methylation on gene expression; relationships in prostate cancer. *Biomolecules*. 2017;7:15.
- Chatterjee R, Vinson C. CpG methylation recruits sequence specific transcription factors essential for tissue specific gene expression. *Biochim Biophys Acta*. 2012;1819:763–770.
- Grossman RL, Heath AP, Ferretti V, et al. Toward a shared vision for cancer genomic data. *N Engl J Med*. 2016;375:1109–1112.
- Zhu Y, Qiu P, Ji Y. TCGA-assembler: open-source software for retrieving and processing TCGA data. *Nat Methods*. 2014;11:599–600.
- Wei L, Jin Z, Yang S, Xu Y, Zhu Y, Ji Y. TCGA-assembler 2: software pipeline for retrieval and processing of TCGA/CPTAC data. *Bioinformatics*. 2017;34:1615–1617.
- R Core Team. R: a language and environment for statistical computing. Vienna, Austria: R Foundation for Statistical Computing. Website: <http://www.R-project.org/>. Up-dated 2015.
- Sarda S, Zeng J, Hunt BG, Yi SV. The evolution of invertebrate gene body methylation. *Mol Biol Evol*. 2012;29:1907–1916.
- Messerschmidt DM, Knowles BB, Solter D. DNA methylation dynamics during epigenetic reprogramming in the germline and preimplantation embryos. *Gene Develop*. 2014;28:812–828.
- Smith ZD, Meissner A. DNA methylation: roles in mammalian development. *Nat Rev Genet*. 2013;14:204–220.
- Mendizabal I, Zeng J, Keller TE, Yi SV. Body-hypomethylated human genes harbor extensive intragenic transcriptional activity and are prone to cancer-associated dysregulation. *Nucleic Acids Res*. 2017;45:4390–4400.
- Huang X, Liu F, Zhu C, et al. Suppression of KIF3B expression inhibits human hepatocellular carcinoma proliferation. *Dig Dis Sci*. 2013;59:795–806.
- Chadalapaka G, Jutooru I, Safe S. Celestrol decreases specificity proteins (Sp) and fibroblast growth factor receptor-3 (FGFR3) in bladder cancer cells. *Carcinogenesis*. 2012;33:886–894.