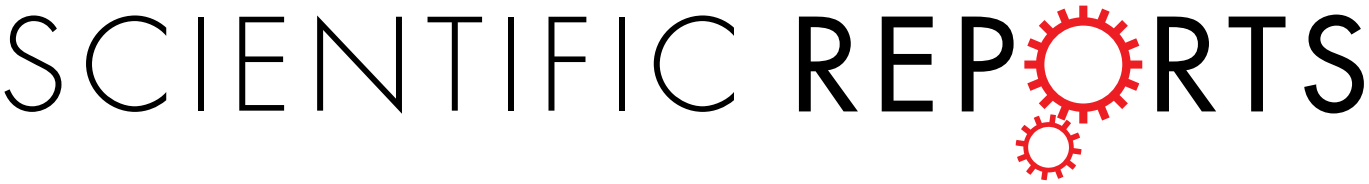

OPEN

Integration of DNA methylation and gene transcription across nineteen cell types reveals cell type-specific and genomic region-dependent regulatory patterns

Received: 26 January 2017
Accepted: 8 May 2017
Published online: 15 June 2017

Binhua Tang[1,2], Yufan Zhou[3], Chiou-Miin Wang[3], Tim H.-M. Huang[3] & Victor X. Jin[3]

Despite numerous studies done on understanding the role of DNA methylation, limited work has focused on systems integration of cell type-specific interplay between DNA methylation and gene transcription. Through a genome-wide analysis of DNA methylation across 19 cell types with T-47D as reference, we identified 106,252 cell type-specific differentially-methylated CpGs categorized into 7,537 differentially (46.6% hyper- and 53.4% hypo-) methylated regions. We found 44% promoter regions and 75% CpG islands were T-47D cell type-specific methylated. Pyrosequencing experiments validated the cell type-specific methylation across three benchmark cell lines. Interestingly, these DMRs overlapped with 1,145 known tumor suppressor genes. We then developed a Bayesian Gaussian Regression model to measure the relationship among DNA methylation, genomic segment distribution, differential gene expression and tumor suppressor gene status. The model uncovered that 3′UTR methylation has much less impact on transcriptional activity than other regions. Integration of DNA methylation and 82 transcription factor binding information across the 19 cell types suggested diverse interplay patterns between the two regulators. Our integrative analysis reveals cell type-specific and genomic region-dependent regulatory patterns and provides a perspective for integrating hundreds of various omics-seq data together.

With the completion of second phase project of the Encyclopedia of DNA elements (ENCODE), thousands of regulatory elements within non-coding regions are now mapped and annotated within the human genome[1–4]. Thus, comprehensive understanding on their roles in mammalian development and human disease progression becomes increasingly important[5,6]. DNA methylation, one of the key epigenetic modifications, plays crucial roles in mammalian cell differentiation, development, and proliferation[5,7,8], and cancer initiation[9–11], such as colorectal cancer[12] and leukemia[13,14] and breast cancer[15–17]. Despite numerous studies done on understanding the role of DNA methylation, there is very limited work on systems integration of DNA methylation with gene expression and transcription factor (TF) binding across multiple cell types at a genome-wide manner.

In this study, we fully utilized the available ENCODE data resource, and conducted genome-wide integration of DNA methylation, TF binding, and RNA expression across 19 cell types. We first computationally identified cell type-specific differentially-methylated regions (DMR), then examined the genomic region specificity for these cell type-specific DMRs. We further developed a Bayesian regression model with Markov Chain Monte Carlo (MCMC) sampling technique to characterize the underlying statistical association among DNA methylation, genomic segment distribution, differential gene expression and tumor suppressor gene (TSG) status. We particularly examined the differential expression of TSGs in these regulatory regions since TSG is involved in many signaling pathways. The identification of TSGs and understanding their relations with DNA methylation

[1]Epigenetics & Function Group, School of the Internet of Things, Hohai University, Jiangsu, 213022, China. [2]School of Public Health and Biostatistics, Shanghai Jiao Tong University, Shanghai, 200025, China. [3]Department of Molecular Medicine, University of Texas Health Science Center, San Antonio, TX, 78229, USA. Binhua Tang and Yufan Zhou contributed equally to this work. Correspondence and requests for materials should be addressed to V.X.J. (email: jinv@uthscsa.edu)

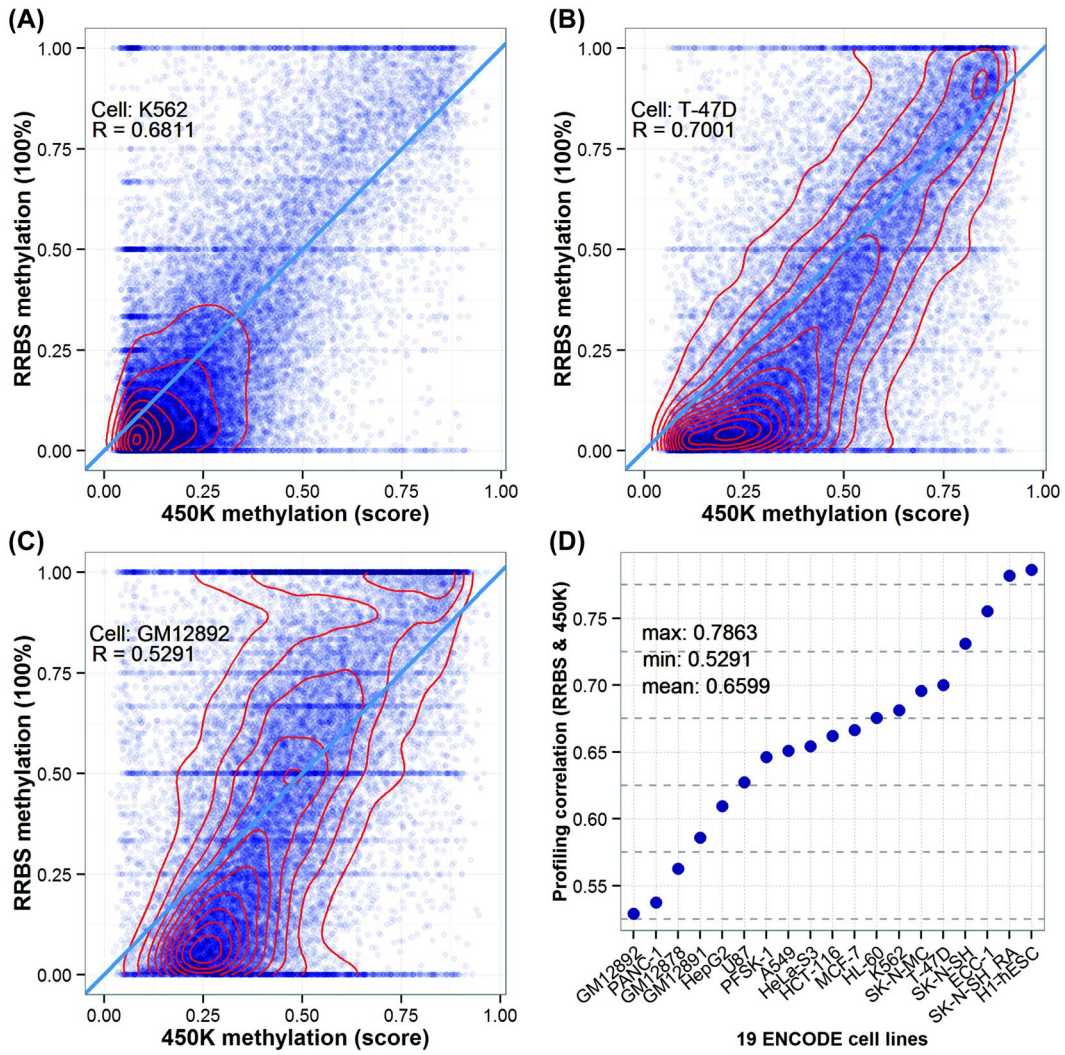

Figure 1. Genome-wide DNA methylation profiling comparison between two major platforms, Illumina Infinium Methylation Beadchip 450 K *vs*. RRBS. (**A**–**C**) Provide the illustrative examples for the three DNA methylation distribution patterns (unimodal, bimodal and trimodal) and corresponding *Pearson* correlation coefficient between two profiling platforms (top left corner); (**D**) presents the correlation statistics between two DNA methylation profiling platforms across the 19 ENCODE cell types.

are critical for further investigation of tumorigenesis[18–20]. We finally applied Kolmogorov-Smirnov (K-S) test in discriminative analysis of the background model and TF-specific methylation pattern[21] to quantitatively examine the interplay of DNA methylation and TF binding.

Results

Examining cell type-specific patterns between array- and sequencing-based DNA methylation profiling. To determine cell type-specific differential DNA methylation across 19 ENCODE cell types (Supplemental Material Table S1), we first examined the quantitative difference between two widely adopted platforms, array-based Illumina Infinium Methylation Beadchip 450 K[22,23], and sequencing-based reduced representation bisulfite sequencing (RRBS)[24,25]. We performed the kernel density estimation on DNA methylation distribution to engender profiling contours and observed a clear unimodal, bimodal and trimodal distribution pattern for K562 (A), T-47D (B) and GM12892 (C) respectively (Fig. 1). Most other 16 cell types followed a bimodal pattern (Supplemental Material Figure S1(I)), except that GM12878 (Figure S1(I) C) and GM12891 (Figure S1(I) D) showed a trimodal distribution pattern and HepG2 (Figure S1(I) M) and PANC-1 (Figure S1(I) O) displayed a divergent distribution in a central section. Overall, the average correlation of two profiling platforms is 0.6599 (Fig. 1D), with a range from 0.7863 (H1-hESC) to 0.5291 (GM12892).

Interestingly, we found 450 K profiling tends to detect lowly-methylated CpG sites due to its probes scattered across the gene, while RRBS can identify highly-methylated CpG sites due to the restriction enzyme used particularly on enriched methylated regions. Our results are useful in determining which platform is suitable for

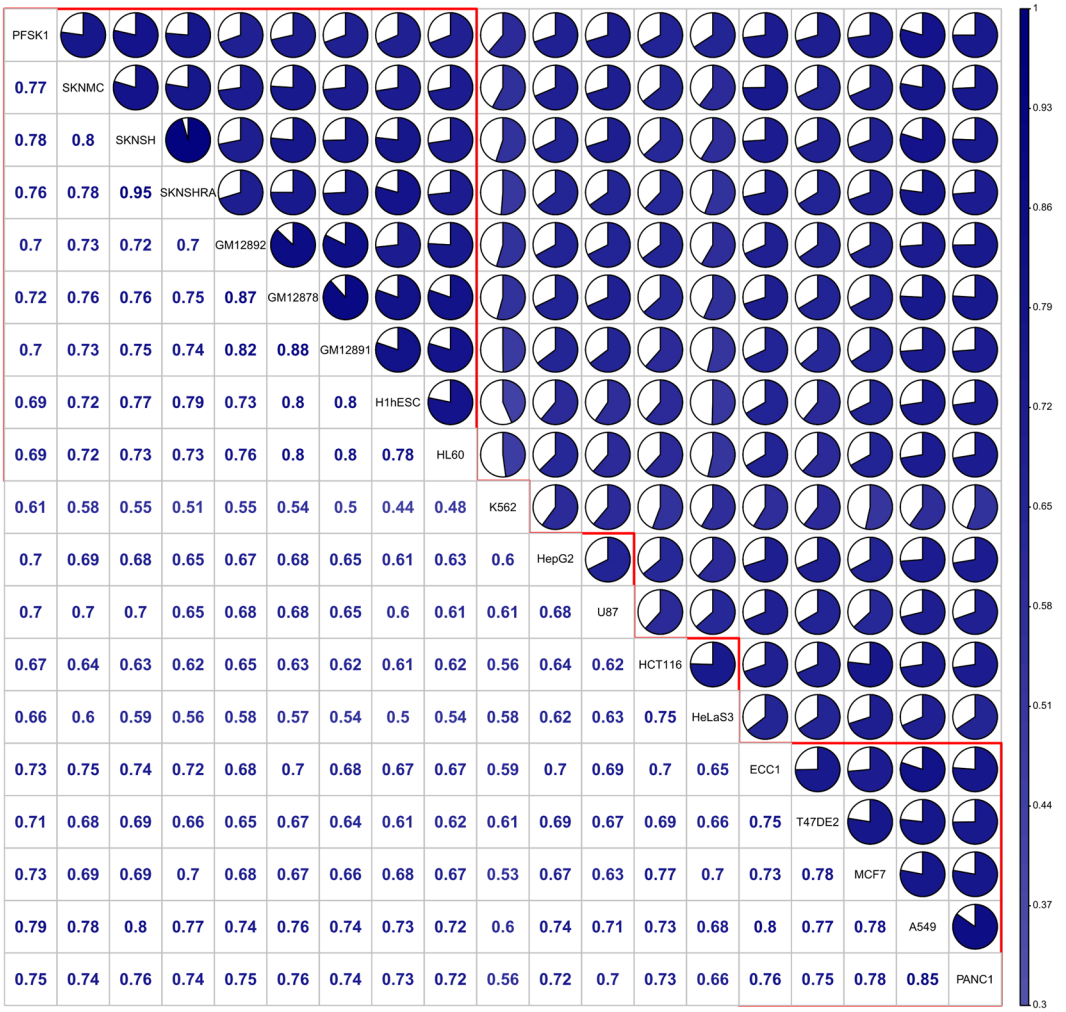

Figure 2. Genome-wide *Pearson* correlation chart for DNA methylation CpG base profiles across 19 cell types. Each diagonal entry gives cell name, upper off-diagonal entry denotes pie chart for the pairwise correlation level and lower off-diagonal entry denotes the corresponding pairwise correlation coefficient for genome-wide methylation of each cell line. The detailed corresponding methylation (Illumina Infinium 450 K) statistics for each cell line was in the Supplemental Material Section 2.

experiment design, suggesting that RRBS is a choice for large-scale samples or population-based studies in despite of a relatively higher cost than 450 K.

We further performed a pairwise correlation among any of two cell types using RRBS data and found an average pairwise correlation coefficient is 0.6908 (Fig. 2). We observed the following correlations: 1) Two breast cancer cell types, MCF-7 and T-47D, have a relatively higher correlation of 0.78; 2) A pancreatic cancer cell type, PANC-1, has higher correlation with some of solid cancer cell types, 0.78 with MCF-7, 0.75 with T-47D, 0.85 with a lung cancer cell type, A549, 0.80 with a endometrial cancer cell type, ECC-1, respectively; 3) Three brain tumor cell types, SK-N-SH, SK-N-MC and SK-N-SH_RA, have noticeably high pairwise correlations, from 0.78 to 0.95, but low correlation with two other brain tumor cell types, PFSK-1 and U87 ($0.65 \leq R \leq 0.78$); 4) Three blood cell types, GM12878, GM12891 and GM12892, have higher correlations with each other ($R > 0.80$) as well as with all other cell types except a leukemia cell type, K562 ($R \sim 0.5$) and a cervical cancer cell type, HeLa-S3 ($0.54 \leq R \leq 0.58$). This may be due to their trimodal methylation distribution patterns; 5) As expected, K562 has relatively lower correlations with all other cell types ($0.44 \leq R \leq 0.61$) partially due to its unimodal pattern (Fig. 1). Our results demonstrated that similar cell origins regardless of their morbidities (cancer or normal) have high correlations for their genome-wide methylation distribution patterns.

The histogram of CpG methylation distribution (RRBS and 450 K) for each of 19 cell types is shown in Supplemental Material Figure S2.

Identifying cell type-specific differentially-methylated regions. Next, we sought to identify the differentially-methylated regions (DMRs) using RRBS data in 19 cell types, and further to interrogate their cell type specificity. Given the fact that 64.24% of CpG sites on average among 19 cell types were covered ≥ 10 sequencing reads, we only used those sites satisfying the sequencing depth with ≥ 10 reads for downstream analyses (see a detail of reads coverage for those cell lines in Supplemental Material Table S1).

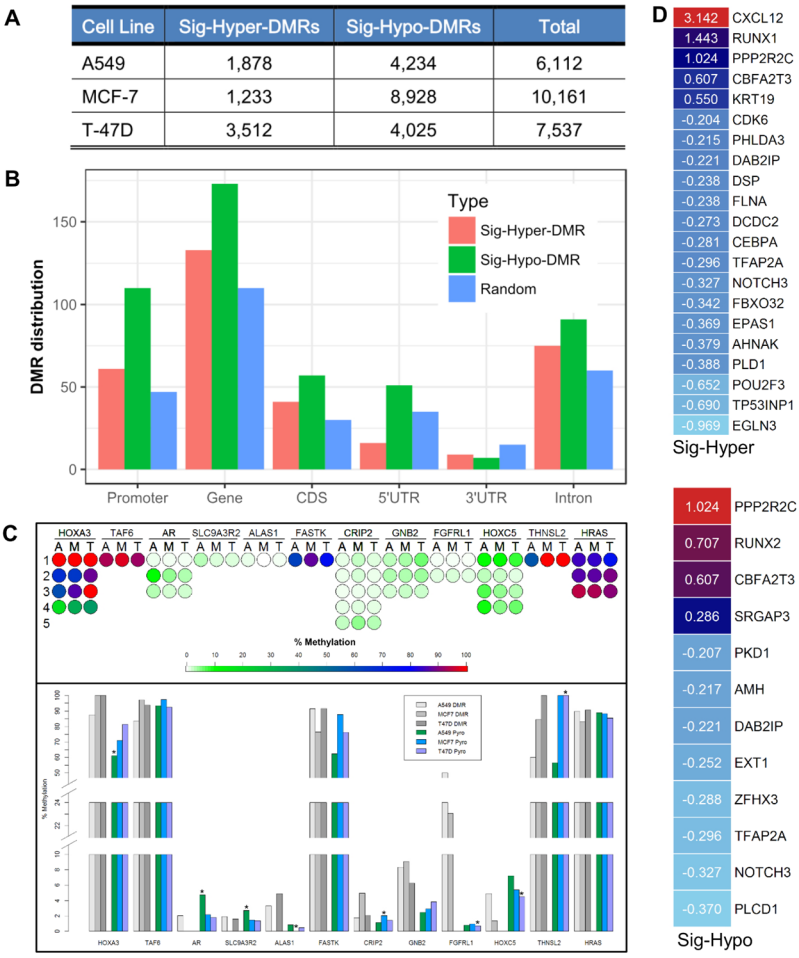

Figure 3. Genome-wide interrogation of significant differentially-methylated CpG sites and regions across 19 cell types. (**A**) The significantly hyper- and hypo-DMRs identified from three benchmark cell types, A549, MCF-7 and T-47D; (**B**) The identified TSGs' components overlapping with Sig-Hyper-DMR and Sig-Hypo-DMR in T-47D; (**C**) Upper panel: DNA methylation pattern of 12 genes in the first 1~5 CpG sites has been detected by pyrosequencing (A: A549; M: MCF7; T: T47D). Lower: average of pyrosequencing detected %Methylation of 1~5 CpG sites for 12 genes in the three cell lines (A549 Pyro: green, MCF7 Pyro: sky blue, T47D Pyro: dark purple) and their DMRs (A549 DMR: light gray, MCF7 DMR: medium gray, T47D DMR: deep gray). Asterisk represents the results as algorithm expected; (**D**) The differential expression status for the investigated TSGs using RNA-seq data (left panel: Sig-Hyper-DMRs; right panel: Sig-Hypo-DMRs).

As a demonstrated cell type, we selected T-47D as a reference since this ERα + breast cancer cell type is used in many studies in our lab and commonly used by other cancer researchers. We detected 688,445 differentially-methylated CpGs (DMCs), and further categorized them into 106,252 significantly DMCs, including 52,232 hyper- and 54,020 hypo-DMCs (threshold of absolute methylation difference $\geq$ 25%, adjusted q-value $\leq$ 0.01). Of them, 32% DMCs locate in promoter regions, 12% in exons, 25% in introns and 32% in intergenic regions; meanwhile, 56% of those DMCs distribute within CpG islands, and 12% at CpG islands shores. Interestingly, we found that promoter regions and CpG islands tend to be hypo-methylated in T-47D, containing 44% and 75% of total hypo-DMCs, respectively; and intergenic regions and CpG island shores only host 25% and 9% of total hypo-DMCs, respectively (Supplemental Material Figure S3). The results indicate that DMCs in promoter or CpG islands are more associated with cell type-specific transcriptional regulation.

We further categorized all statistically significant DMCs into DMRs with the published toolkits[26–29]. We preprocessed by methylKit with the thresholds: region's mean methylation difference cutoff $\geq$ 20% (adjusted q-value $\leq$ 0.01), and we obtained 16,277 DMR candidates. Of all the identified DMRs in T-47D cells, 8,936 are hyper-methylated (Hyper-DMR) and 7,341 are hypo-methylated (Hypo-DMR). With the more stringent adjusted q-value $\leq$ 0.001 and differentially-methylated CpG base count $\geq$ 5 to define DMR, we detected 7,537 statistically significant DMRs (Sig-DMR), 3,512 significant hyper-DMRs (Sig-Hyper-DMR) and 4,025 significant hypo-DMRs (Sig-Hypo-DMR). We also detected MCF-7 and A549 cell type specific DMRs respectively (Fig. 3A).

We associated those T-47D specific DMRs with ~1,200 currently known tumor suppressor genes[18], by overlapping each DMR with TSGs' genomic regions, promoter, 5′UTR, 3′UTR, CDS, and intron. We found 133 TSGs overlapping with the Sig-Hyper-DMRs and 173 TSGs with Sig-Hypo-DMRs, respectively (Fig. 3B).

Together we experimentally validated the cell type-specific DMRs in the selected three cell types, A549, T-47D and MCF-7 with a pyrosequencing assay (Fig. 3C). Our results illustrated that DNA methylation of the first 1~5 CpG sites in the randomly selected 12 genes has been detected (A: A549; M: MCF-7; T: T-47D, Fig. 3B - upper panel). An average of methylation of 1~5 CpG sites for eight genes showed their cell type-specific methylation patterns in A549 (red), M-CF7 (blue) and T-47D (green) (Fig. 3C - lower panel), respectively. The other four genes only showed cell type-specific methylation at individual CpG sites.

Furthermore, we examined their differential expression levels in T-47D with estrogen (E2) treatment vs. control (DMSO) samples using RNA-seq data from ENCODE, and found many genes, especially the corresponding TSGs, showed significantly differentially expressed (absolute log2 fold change ≥ 0.2, adjusted p-value ≤ 0.05; Fig. 3D, upper panel depicting the differential expression of TSGs in Sig-Hyper-DMRs and lower panel for TSGs in Sig-Hypo-DMRs). The Supplemental Material Figure S4 presents the genome-wide DMC distribution and related statistical properties.

Interrogating genomic region-dependent DNA methylation patterns. During cancer development and progression, changes in DNA methylation occur mostly within gene promoter (2,000 bp centered on TSS), CpG island and CpG island shore (1,000 bp centered around CpG island)[11,30]. Thus, we particularly examined six genomic regions to understand genomic region-dependent DNA methylation patterns. In addition, we were only focused on these regions with TF binding. Region I – CGI.NP: CpG islands outside promoter regions, II – CGIS. NP: CpG island shores outside promoter regions, III – P.CGI: CpG islands within promoter regions, IV – P.CGIS: CpG island shores within promoter regions, V – P.NCGI: promoter regions out of CpG islands and VI – P.NCGIS: promoter regions out of CpG island shores, illustrated in the Supplemental Material Figure S5(I).

We observed there exists two distinct region-dependent patterns. One is hyper-methylation in Regions I, II, III and even in IV, but semi-hyper-methylation in Regions V and VI (Fig. 4A–C), and the other is hypo-methylation in Regions I, II, III, and IV, and even low methylation in Regions V and VI (Fig. 4D–F). The results indicated that DNA methylation in CpG islands and their shores exerted a cell type specificity than those in promoters since Regions V and VI (the promoter sub-regions independent of CpG islands and their shores) generally showed hypo-methylated. These distribution patterns were recapitulated in many TFs from other cell types (Supplemental Material Figure S5).

We speculated that such features, especially TFs' binding at the hypo-methylated regions, might be critical for studying transcriptional regulation in cancer cells, such as ERα and CTCF in T-47D cell, as hypo-methylation triggers targeted promoter activity and its downstream transcriptional regulation[8,31]. Indeed, we also found the CpG island-related sub-regions (independent of promoters) were overlapped with several sets of known enhancers. This is in line with other previous findings that enhancers, rather than promoters, are embedded comparatively more various DMRs in the context of a specific developmental and disease stage[32,33]. It is noted that in addition to enhancer regions there might be other non-coding regions within the CGI-related sub-regions that might be important for transcriptional activities.

Statistically integrating DNA methylation and gene expression. Next, we proposed a Bayesian regression model using Markov Chain Monte Carlo method to characterize the underlying association among DNA methylation, genomic segment distribution, differential gene expression level and tumor suppressor gene status (See the Method section for the details). The MCMC sampling distributions for modeling were given in Fig. 5, with the trace from genomic segment 3′UTR omitted automatically due to the insufficient sample size acquired, 25 against total 1,590 for the hyper case, and 38 against 1,291 for the hypo case, respectively. The average sample size for the other five genomic segments is 563. Figure 5 lists the regression coefficient median (med) by MCMC sampling, together with 95% confidence interval (CI) indicated by a brown line below, with the range at each end.

To statistically validate the derived model, we further introduced other regression item, HPRO (hyper- or hypo-methylation), into the model; and compared both log marginal likelihood and selection possibility for the models using Bayes Factor method. We found that Equation 2 is statistically fit for the current data with much higher selection possibility, 0.9855578 against 0.0144422 for the model with another item HPRO introduced, although the log marginal likelihoods are similar for two models, −1335.051 for the initially proposed model and −1339.274 for the model with a new item. This difference is due to duplicate effects from the variants Methy and HPRO. Methy takes positive or negative differential methylation value, which has quantitatively reflected binary hyper- or hypo-methylation status, HPRO.

Thus, the model in Equation 2 can best depict the quantitative relationship of differential gene expression level for TSG with DNA methylation, and genomic segment distribution. The Supplemental Material Figures S6–S8 illustrated the hyper-, hypo-methylation cases, and the case for the model with a new variant HPRO introduced.

Unraveling the interplay between DNA methylation and TF binding. To determine the quantitative interplay between DNA methylation and 82 TFs binding across the 19 cell types (Fig. 6), we introduced a Beta-distribution model for fitting the background methylation profile of each of 82 TFs in regions of a 2,000 bp length centered on TSS (TSS $\pm$ 1000 bp).

We then utilized a Kolmogorov-Smirnov (K-S) test in iteratively discriminating the background methylation status from TF-specific methylation pattern[21] (Fig. 6A,B).

Based on K-S discriminative test analysis together with multiple testing correction, we identified 157 highly-differentiated methylation-TF (BH-adjusted p-value ≤ 0.05), 156 lowly-differentiated methylation-TF, namely the lowly-differentiated methylation-TF (BH-adjusted p-value > 0.05), and 1,327 un-differentiated methylation-TF (gray block), see Fig. 6C.

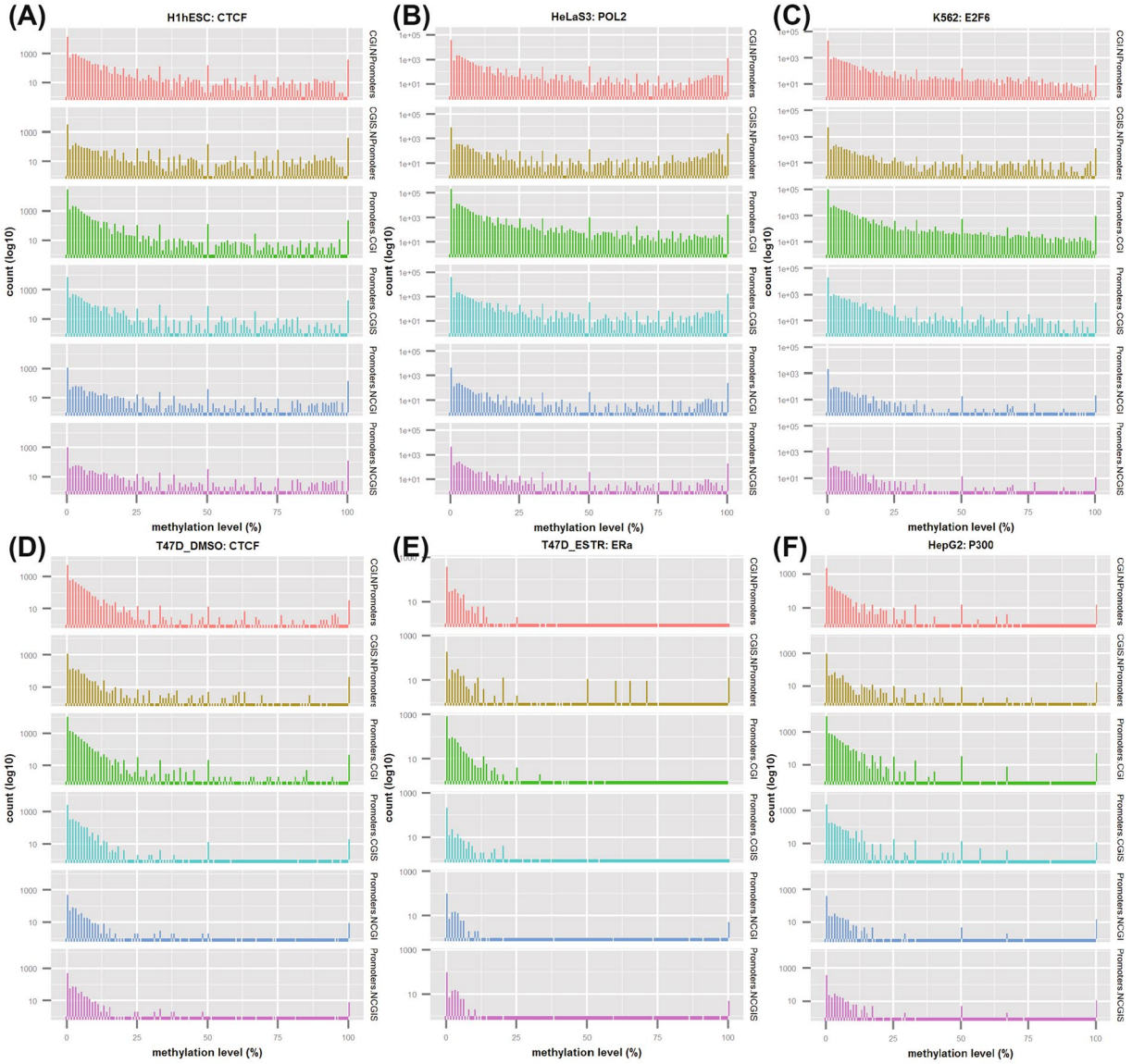

Figure 4. Interrogation of cell type-specific and genomic region-dependent DNA methylation distribution among six genomic regions across five cell types, namely H1-hESC, HeLa-S3, K562, T-47D (DMSO), T-47D (ESTR) and HepG2, within the cell-type and TF combinations. X-axis depicts the methylation level distribution (0~100%) and y-axis for the CpG loci (sites) count at each specific methylation level.

Our results showed many TF binding patterns are irrespective to their methylation status, while some key TFs prefer binding at lowly-methylated promoters. For example, POL2 (RNA polymerase II) showing highly-differential methylation among 10 cell types is consistent with its role as a form of eukaryotic RNA polymerase II, recruiting other core factors to the promoters of protein-coding genes during transcription initiation. ERα (estrogen receptor α) is highly differentially-methylated in ECC-1 cell; CEBP (an enhancer-binding protein) and TCF12 are found to be highly differentiated in MCF-7 cell, and interestingly, P300 (a transcription coactivator) and CTCF (an insulator) are also identified as highly differentially-methylated entries[34–37]. Our results suggest a diversified interplay mechanism between DNA methylation and TF binding activity across cell types.

Material and Methods

DNA methylation RRBS and ChIP-seq data sources. We obtained reduced representation bisulfite sequencing (RRBS) data and ChIP-seq data of 82 TFs in 19 cell types from the ENCODE Consortium Project[1], which cover major human cancer cell types, human blood B-lymphocyte and embryonic stem cell types (Supplemental Material Table S1). We also retrieved Illumina Infinium Methylation Beadchip 450 K data, a CpG-specific array technology profiling over 450,000 CpGs covering 99% of all RefSeq genes, to perform comparison and correlation analysis with the RRBS data[23,38,39].

Definition of cell type-specific differentially-methylated CpGs and regions. We inferred the cell type-specific DMRs with a reference to a particular cell type based on the following definition:

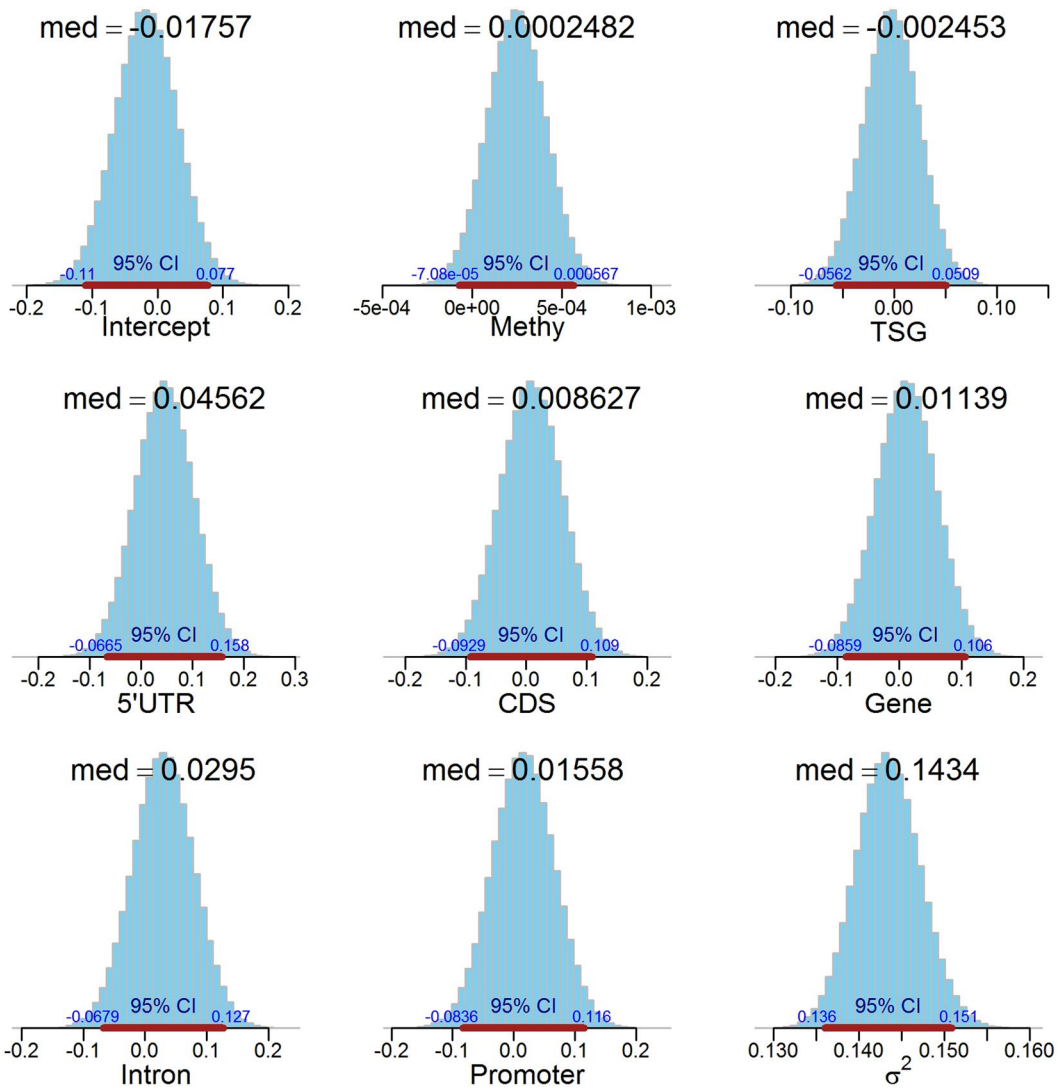

Figure 5. Statistics for Gaussian regression for differential expression (log2 fold change from RNA-seq) based on methylation level, TSG status, and DMR annotation region distribution (gene body, intron, promoter, 5′UTR, 3′UTR and CDS). Each plot lists the median (med) for sampling distribution, together with 95% confidence interval (CI) and start/end positions.

$$\widehat{DMR_i} = DMR_i \setminus \bigcup_{j \in S \setminus i} DMR_j \tag{1}$$

where DMR_i denotes the DMR set identified for the i-th cell type from the 19 cell type set S, $\bigcup_{j \in S \setminus i} DMR_j$ denotes the DMR set union by all the other cell types from S except for the i-th cell type, the symbol\denotes the set deduction operation, and $\widehat{DMR_i}$ indicates cell type-specific DMR set for the i-th cell type after set subtraction from the other DMR sets.

Biological validation experiments with pyrosequencing assay. Genomic DNA of MCF-7, T-47D and A549 cell lines were extracted with QIAamp DNA Mini Kit (QIAGEN) respectively, then bisulfite-converted with EZ DNA Methylation-Gold Kit (ZYMO RESEARCH). The DNA sequences of 12 cell type-specific DMRs were obtained from human genome, then input to Qiagen PyroMark Assay Design 2.0 software. PCR primers and pyrosequencing primers for the pyrosequencing experiments were generated by the software. Genes were amplified with primers listed in Table 1 using the converted genomic DNA as template.

The PCR products were verified for size and quality with agarose gel electrophoresis. Biotin-labelled forward or reverse primers were used for PCR and the products were pulled down with streptavidin-coated agarose beads. The biotinylated single-strand DNA fragments were purified with PyroMark Q96 Vacuum Workstation (QIAGEN). The pyrosequencing was performed with pyrosequencing primers listed in Table 1 using PyroMark Q96 MD (QIAGEN). The sequencing results were analyzed with PyroMark CpG software (QIAGEN) to get the percentage of methylation level (%Methylation).

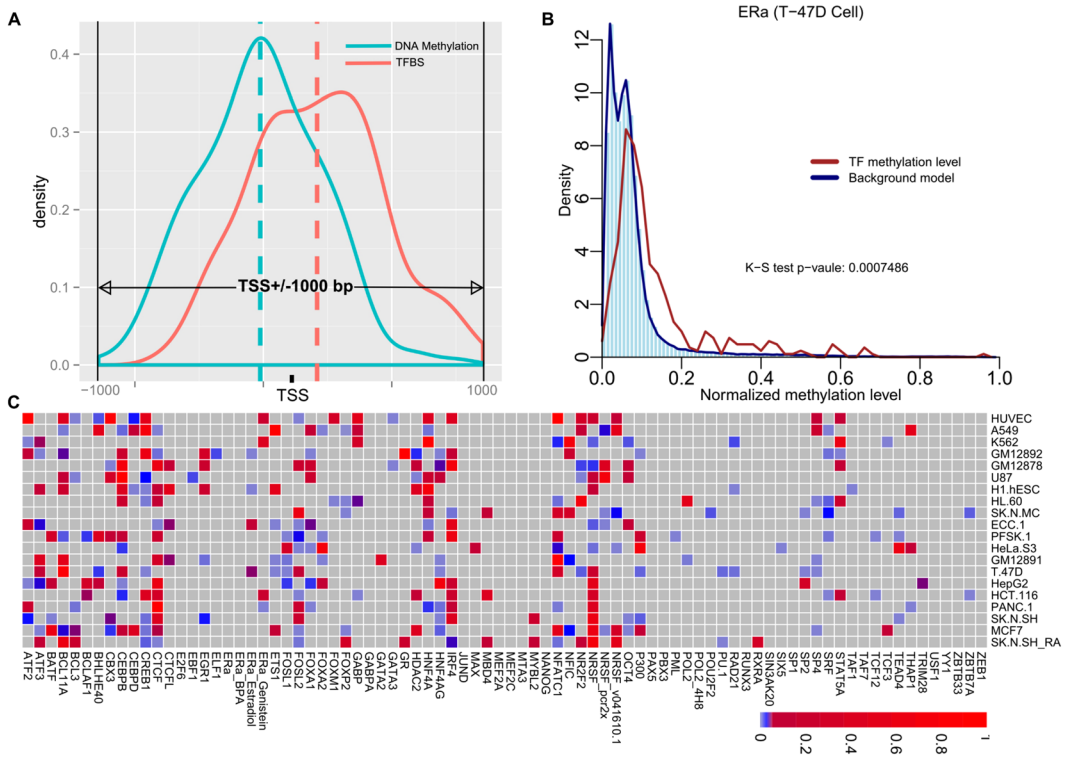

Figure 6. Genome-wide integration of DNA methylation and 82 TF binding from 19 cell types. (**A**) Schematic illustration in detecting differential DNA methylation status at TFBS regions (TSS ± 1000 bp) by Kolmogorov-Smirnov test; (**B**) Illustrative example for ERα in T-47D cell, blue dashed curve and histogram represent the background model in density distribution plot, the brown curve denotes the density plot for DNA methylation distribution at ERα binding site with range from 0 to 1; (**C**) A heatmap showing the differentially-methylated TFs, where vertical axis lists the 19 cell types and horizontal axis for the 82 TFs, 157 highly-differentiated entries (K-S test against the background model, BH-adjusted p-value ≤ 0.05, light blue to blue color), 156 lowly-differentiated entries (K-S test against the background model, BH-adjusted p-value > 0.05, red color), and 1,327 null entries (gray color).

Statistical modeling the association between DNA methylation and gene expression. We developed a Bayesian regression model with Markov Chain Monte Carlo sampling method to characterize the statistical association among differential gene expression (DGE) level (log2 fold change) acquired from the control and treatment samples by RNA-seq, methylation level (Methy), TSG category, and methylation level regarding genomic segment (GS) distribution. The Bayesian regression model takes the following form,

$$\begin{aligned} \mathrm{DGE}_i &= XB + \Xi_i \\ &= \sum_{i=1,j=1}^{m,n} x_{i,j}\beta_{i,j} + \varepsilon_i \\ &= \sum_{i=1}^{m}[\mathrm{Methy}_i\beta_{1,i} + \mathrm{TSG}_i\beta_{2,i} + \mathrm{GS}_i\beta_{3,i}] + \varepsilon_i \end{aligned} \tag{2}$$

where $i \in N$, DGE_i stands for log2 fold change values for i-th annotated gene, Methy_i stands for a mean methylation value (in percentage) of the i-th annotated gene's hosted DMR, TSG_i for the status whether or not the i-th annotated gene belongs to TSG (binary TRUE or FALSE), GS_i for the i-th gene's annotated genomic segment base (promoter, 5′UTR, gene body, introns and 3′UTR), and the error entry follows Gaussian distribution, $\varepsilon_i \sim N(0, \sigma^2)$.

We further assume standard semi-conjugate priors,

$$\begin{aligned} \beta &\sim N(b_0, B_0^{-1}) \\ \sigma^{-2} &\sim Gamma(c_0/2, d_0/2) \end{aligned} \tag{3}$$

where β and σ^{-2} are assumed a priori independent, b_0, B_0^{-1}, c_0 and d_0 are the initial values predefined for Gaussian and Gamma distribution, respectively.

Bayes Factor analysis was adopted to compare the model candidates and perform model selection using log marginal likelihood and selection possibility. Then the posterior information of regression model with Gaussian

Gene	PCR Forward	PCR Reverse	Pyrosequencing Primer	Biotin-labelled
HOXA3	GTTTAGGGAAGGGTTGGTTTAG	AACAATCCAACTCCTAATCTCCTCC	GGGTGATTTTTTTTAGTTTAGT	Reverse
TAF6	GGTTAAGGTGGTAGTTTGTT	CTAATCTTAAAACTAAAAACTCTTC	AACTCTTCCTACCCA	Forward
AR	AAGGAGGTGGGAAGGTAAG	ACTAACTCCACCCTTTTTCCCTCTATC	GTTGTATTTGTTTTTTATTTTTTAG	Reverse
SLC9A3R2	TAGAGTAGGGGAGAGATAGAGAGGTT	ACCAAACCCCTACCT	CCAAACCCCTACCTC	Forward
ALAS1	TAGGATGAGGTAGGGAAAAGAGATT	AAAAAACCAAACAAAAAAACCCACTTCTTA	AGGAGAGTAGGGGAATTT	Reverse
FASTK	AATGGAAGAGGAGGGGGATTTAGT	ACCCCCCCATAAAAAATAAATAATTTACAC	TTGTTTTTAGTTTTAAAATTGAGAT	Reverse
CRIP2	GTTATGGAAATAGAGTAATAAAGGGAAG	AACCCTAATAACTTAAACCTAAAAATCC	GGAGTTTGAGATTTTTTT	Reverse
GNB2	GTGGGAGAGGTTGAGGAAATGTT	CCACCCCCCTCACCAAA	AAAACCAAACTAAAAAAACCTAAAC	Forward
FGFRL1	GGGTTAGGGGTTTTAGTTGG	ACCCCCAAAACACACAACACTCAA	GGGGTTTTAGTTGGGTA	Reverse
HOXC5	TTTTATGAGAGAATTGGGTAAATATGG	TAACTTCTTATAACCAATCCAACTTA	ACTATAAATTTTCTACAAACAACC	Forward
THNSL2	ATGTGTTTAGGAGATTGGTGGTTTAGA	CCCCTAATACTAACATACCACATAATCCT	GTGTGTGTTTTGGTTT	Reverse
HRAS	AGAGTTTAGGGTTGGATAGGT	CTCCCAAACCTCTATAAACCCTTATC	GTTTGAGGTTTAGATATATTTTATG	Reverse

Table 1. Primers used for pyrosequencing experiment.

errors is acquired using Gibbs sampling[40–43]. We sampled 500,000 iterations with the first 1,000 times truncated to ensure trace convergence.

Detecting differential DNA methylation status at promoter regions. To identify TFs with specific methylation status around their TSS regions, we adopted Kolmogorov-Smirnov (K-S) test for statistically differentiate the background model and each TF's DNA methylation distribution. For each TF, its peak binding region for measuring methylation level is defined as a length of 2000 bp centered on TSS (TSS $\pm$ 1000 bp), mainly for covering each promoter region.

We selected the RRBS data for measuring each TF's methylation status. With the aggregated DNA methylation level from the 82 TFs across all the 19 cell types, we constructed their methylation distribution background model with the normalization and fitting. Then we adopted the K-S test for determining the statistic p-value between TF's methylation distribution density and the background model. We defined those entries with p-value ≤ 0.05 as the significantly differential methylated compared with background model.

Tools used in the curation and analysis. Bowtie2[44] was used to align sequencing reads, SAMtools[45] and BAMtools[46] were used to process the aligned sequencing reads, methylKit[29] was used to analyze part of RRBS data, and DESeq[47] was used to analyze RNA-seq data.

Discussion

Integration of multi-platform and cross-cell-types omics information enables the thorough interrogation of genomic features with undiscovered biological functions. Till now, there is still limited systematic analysis for hundreds of data sets across different data types and multiple cell types.

Our work conducted the systems integration of 19 ENCODE cell types about the cell type-specific DNA methylation and its impacts on transcriptional regulation. The systematic analysis on DNA methylation within predefined genomic regions or segments revealed that DNA methylation in CpG islands and CpG islands shore characterizes a specific cell type's methylation status, which may act as a hallmark in studying transcriptional regulation for the different cancer cells.

We found that promoter regions and CpG islands in T-47D tend to be hypo-methylated, with 44% and 75% of total significantly hypo-methylated DMCs, respectively while CpG islands shores only cover 9%. Annotation showed hyper- and hypo-methylated regions are embedded with 634 and 511 TSGs, respectively, strongly suggesting the mechanism and functionality underlying the TSGs tightly linked to DNA methylation. Future experiments may be needed to determine whether DNA methylation plays a casual role in hyper-methylation associated TSGs by switching methylation status to hypo-methylation.

Through the cross cell-type comparison of differential gene expression, DNA methylation, and differentially-methylated CpG sites and genomic regions, we derived a quantitative formula to measure their relationship using our proposed Bayesian regression model. Our Bayesian model reveals that methylation sites in 3′UTR have much less impact on transcriptional regulation than other regions.

In all, our systematic analysis reveals cell type-specific and genomic region-dependent regulatory patterns with a breast cancer cell T-47D as a benchmark cell type, and provides an efficient approach in integrating hundreds of various omics-seq data together.

Availability. The analyzed intermediate results and tables for the project are deposited at: https://github.com/gladex/PanCanMAP.

References

1. The ENCODE Project Consortium. An integrated encyclopedia of DNA elements in the human genome. *Nature* **489**, 57–74 (2012).
2. Pennisi, E. ENCODE Project Writes Eulogy for Junk DNA. *Science* **337**, 1159–1161 (2012).
3. de Souza, N. Genomics: The ENCODE project. *Nat Meth* **9**, 1046–1046 (2012).
4. The Encode Project Consortium. A User's Guide to the Encyclopedia of DNA Elements (ENCODE). *PLoS Biol* **9**, e1001046 (2011).
5. Smith, Z. D. & Meissner, A. DNA methylation: roles in mammalian development. *Nat Rev Genet* **14**, 204–220 (2013).

6. Lister, R. *et al*. Global epigenomic reconfiguration during mammalian brain development. *Science* **341** (2013).
7. Reik, W., Dean, W. & Walter, J. Epigenetic reprogramming in mammalian development. *Science* **293**, 1089–1093 (2001).
8. Varley, K. E. *et al*. Dynamic DNA methylation across diverse human cell lines and tissues. *Genome Research* **23**, 555–567 (2013).
9. Das, P. M. & Singal, R. DNA Methylation and Cancer. *Journal of Clinical Oncology* **22**, 4632–4642 (2004).
10. Aran, D. & Hellman, A. DNA methylation of transcriptional enhancers and cancer predisposition. *Cell* **154**, 11–13 (2013).
11. Witte, T., Plass, C. & Gerhauser, C. Pan-cancer patterns of DNA methylation. *Genome Medicine* **6**, 1–18 (2014).
12. Irizarry, R. A. *et al*. The human colon cancer methylome shows similar hypo- and hypermethylation at conserved tissue-specific CpG island shores. *Nat Genet* **41**, 178–186 (2009).
13. Xu, J. *et al*. DNMT3A Arg882 mutation drives chronic myelomonocytic leukemia through disturbing gene expression/DNA methylation in hematopoietic cells. *PNAS* **111**, 2620–2625 (2014).
14. Nordlund, J. *et al*. Genome-wide signatures of differential DNA methylation in pediatric acute lymphoblastic leukemia. *Genome Biology* **14**, r105 (2013).
15. Szyf, M. DNA methylation signatures for breast cancer classification and prognosis. *Genome Medicine* **4**, 26 (2012).
16. Aure, M. *et al*. Individual and combined effects of DNA methylation and copy number alterations on miRNA expression in breast tumors. *Genome Biology* **14**, R126 (2013).
17. Heyn, H. *et al*. DNA methylation profiling in breast cancer discordant identical twins identifies DOK7 as novel epigenetic biomarker. *Carcinogenesis* **34**, 102–108 (2013).
18. Zhao, M., Sun, J. & Zhao, Z. TSGene: a web resource for tumor suppressor genes. *Nucleic Acids Research* **41**, D970–D976 (2013).
19. Osborne, C., Wilson, P. & Tripathy, D. Oncogenes and Tumor Suppressor Genes in Breast Cancer: Potential Diagnostic and Therapeutic Applications. *The Oncologist* **9**, 361–377 (2004).
20. Pradeepkiran, J. A. *et al*. CGMD: An integrated database of cancer genes and markers. *Scientific Reports* **5**, 12035 (2015).
21. Conover, W. J. Pra*ctical nonparametric statistics*. (Wiley, 1999).
22. Houseman, E. *et al*. DNA methylation arrays as surrogate measures of cell mixture distribution. *BMC Bioinformatics* **13**, 86 (2012).
23. Maksimovic, J., Gordon, L. & Oshlack, A. SWAN: Subset-quantile Within Array Normalization for Illumina Infinium HumanMethylation450 BeadChips. *Genome Biology* **13**, R44 (2012).
24. Harris, R. A. *et al*. Comparison of sequencing-based methods to profile DNA methylation and identification of monoallelic epigenetic modifications. *Nat Biotech* **28**, 1097–1105 (2010).
25. Bock, C. Analysing and interpreting DNA methylation data. *Nat Rev Genet* **13**, 705–719 (2012).
26. Zhang, Y. *et al*. QDMR: a quantitative method for identification of differentially methylated regions by entropy. *Nucleic Acids Research* **39**, e58–e58, doi:10.1093/nar/gkr053 (2011).
27. Hebestreit, K., Dugas, M. & Klein, H.-U. Detection of significantly differentially methylated regions in targeted bisulfite sequencing data. *Bioinformatics* **29**, 1647–1653, doi:10.1093/bioinformatics/btt263 (2013).
28. Hansen, K., Langmead, B. & Irizarry, R. BSmooth: from whole genome bisulfite sequencing reads to differentially methylated regions. *Genome Biology* **13**, R83 (2012).
29. Akalin, A. *et al*. methylKit: a comprehensive R package for the analysis of genome-wide DNA methylation profiles. *Genome Biology* **13**, R87 (2012).
30. Stefansson, O. A. *et al*. A DNA methylation-based definition of biologically distinct breast cancer subtypes. *Molecular Oncology* (2014).
31. Liu, H. *et al*. DNA methylation dynamics: identification and functional annotation. *Brief Funct Genomics* **15**, 470–484, doi:10.1093/bfgp/elw029 (2016).
32. Elliott, G. *et al*. Intermediate DNA methylation is a conserved signature of genome regulation. *Nature Communications* **6**, 6363 (2015).
33. Gu, J. *et al*. Mapping of Variable DNA Methylation across Multiple Cell Types Defines a Dynamic Regulatory Landscape of the Human Genome. *G3: Genes|Genomes|Genetics* **6**, 973–986, doi:10.1534/g3.115.025437 (2016).
34. Tang, B. *et al*. Hierarchical modularity in ERα transcriptional network is associated with distinct functions and implicates clinical outcomes. *Scientific Reports* **2** (2012).
35. Chen, Y. *et al*. DNA Binding by GATA Transcription Factor Suggests Mechanisms of DNA Looping and Long-Range Gene Regulation. *Cell Reports* **2**, 1197–1206 (2012).
36. Kemp, C. J. *et al*. CTCF haploinsufficiency destabilizes DNA methylation and predisposes to cancer. *Cell Reports* **7**, 1020–1029 (2014).
37. Handoko, L. *et al*. CTCF-mediated functional chromatin interactome in pluripotent cells. *Nat Genet* **43**, 630–638 (2011).
38. Lister, R. *et al*. Human DNA methylomes at base resolution show widespread epigenomic differences. *Nature* **462**, 315–322 (2009).
39. Bibikova, M. *et al*. High density DNA methylation array with single CpG site resolution. *Genomics* **98**, 288–295 (2011).
40. Martin, A. D., Quinn, K. M. & Park, J. H. MCMCpack: Markov Chain Monte Carlo in R. *Journal of Statistical Software* **42**, 21 (2011).
41. Chib, S. Marginal Likelihood from the Gibbs Output. *Journal of the American Statistical Association* **90**, 1313–1321 (1995).
42. Lin, L.-A., Luo, S., Chen, B. E. & Davis, B. R. Bayesian analysis of multi-type recurrent events and dependent termination with nonparametric covariate functions. *Statistical Methods in Medical Research* (2015).
43. Liu, J. S. *Monte Carlo Strategies in Scientific Computing*. (Springer-Verlag New York, Inc.).
44. Langmead, B. & Salzberg, S. L. Fast gapped-read alignment with Bowtie 2. *Nat Meth* **9**, 357–359 (2012).
45. Li, H. *et al*. The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25**, 2078–2079 (2009).
46. Barnett, D., Garrison, E., Quinlan, A., Strömberg, M. & Marth, G. BamTools: a C++ API and toolkit for analyzing and managing BAM files. *Bioinformatics* **27**, 1691–1692 (2011).
47. Anders, S. & Huber, W. Differential expression analysis for sequence count data. *Genome Biology* **11**, R106 (2010).

Acknowledgements

This work was supported by the Natural Science Foundation of Jiangsu, China (BE2016655 and BK20161196), Fundamental Research Funds for China Central Universities (2016B08914) and Changzhou Science & Technology Program (CE20155050). This work was also partially supported by NIH 1R01GM114142, 1U54CA217297 and the Bioanalytics and Single-Cell Core at UTHSCSA, supported by UTHSCSA and CPRIT grant (RP150600). This work made use of the resources supported by the NSFC-Guangdong Mutual Funds for Super Computing Program (2nd Phase), and the Open Cloud Consortium sponsored project resource, supported in part by grants from Gordon and Betty Moore Foundation and the National Science Foundation (USA) and major contributions from OCC members.

Author Contributions

B.H.T. conceived the study, carried out the design of the pan-cancer study, performed the statistical and computational analysis; Y.Z., C.M.W. and V.X.J. conceived and performed the experiment validation; B.H.T., Y.Z., T.H.H. and V.X.J. drafted the manuscript, read and approved the final manuscript.

Additional Information

Supplementary information accompanies this paper at doi:10.1038/s41598-017-03837-z

Competing Interests: The authors declare that they have no competing interests.

Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.