Hindawi Publishing Corporation
BioMed Research International
Volume 2016, Article ID 1597489, 10 pages
http://dx.doi.org/10.1155/2016/1597489

Research Article

META2: Intercellular DNA Methylation Pairwise Annotation and Integrative Analysis

Binhua Tang[1,2]

[1]*Epigenetics & Function Group, School of Internet of Things, Hohai University, Jiangsu 213022, China*
[2]*School of Public Health, Shanghai Jiao Tong University, Shanghai 200025, China*

Correspondence should be addressed to Binhua Tang; bh.tang@outlook.com

Received 17 September 2016; Accepted 12 December 2016

Academic Editor: Hao-Teng Chang

Genome-wide deciphering intercellular differential DNA methylation as well as its roles in transcriptional regulation remains elusive in cancer epigenetics. Here we developed a toolkit META2 for DNA methylation annotation and analysis, which aims to perform integrative analysis on differentially methylated loci and regions through deep mining and statistical comparison methods. META2 contains multiple versatile functions for investigating and annotating DNA methylation profiles. Benchmarked with T-47D cell, we interrogated the association within differentially methylated CpG (DMC) and region (DMR) candidate count and region length and identified major transition zones as clues for inferring statistically significant DMRs; together we validated those DMRs with the functional annotation. Thus META2 can provide a comprehensive analysis approach for epigenetic research and clinical study.

1. Introduction

Genome-wide DNA methylation analysis and annotation across multiple samples are essential in interrogating pairwise base-pair differences, while it still remains elusive in recent pancancer studies [1–7]. Pancancer DNA methylation study can retrieve cell- and tissue-specific properties by detecting differentially methylated loci and regions.

Heyn et al. adopted the Illumina Infinium 450 K technique to identify DOK7 as novel biomarker in breast cancer [8]; and the genome-wide composition, patterning, cell specificity, and dynamics of DNA methylation at single-base resolution in human and mouse frontal cortex throughout their lifespan were reported recently [9]; Bell et al. applied whole-blood DNA methylation to investigate molecular clues in chromic pain [10].

However, till now, our knowledge about the genome-wide distribution of DNA methylation, how to decipher the genome-wide difference, and how it relates to other epigenetic modifications in mammals remains limited. And there still lacks comprehensive analysis toolkits for biochemical experiment design and postexperiment validation.

Herein we developed an analysis toolkit, META2, for intercellular DNA methylation annotation and analysis. META2 is mainly designed for analyzing the reduced representation bisulfite sequencing (RRBS) profiling data [11–13]; together it can analyze data with the right formats from other platforms, such as HumanMethylation 450 K beadchip assay [14–16]. META2 can implement intercellular interrogation of DNA methylation status among multiple samples, perform statistical analysis on methylated CpG loci and regions, and yield integrative visualization for the analysis results.

We also validated the toolkit on the real RRBS data retrieved from ENCODE consortium and demonstrated its integrative analysis on the last section. Our developed toolkit aims to provide a versatile analysis approach to the epigenetic research fields, and we also deposited the toolkit on GitHub for public convenient usage.

2. Structure and Function Composed in META2

The toolkit META2 contains several major functional procedures, namely, (i) DNA methylation raw data acquisition and

preprocess; (ii) statistical analysis and information retrieval; and (iii) integrative analysis and visualization, as depicted in Figure 1.

The first functional procedure of META2 is the acquisition and curation of raw DNA methylation data, for example, sequencing-based RRBS and array-based 450 K platforms [17, 18]. This procedure covers preprocessing the raw DNA methylation information, from integration of sample list (pairwise control versus treatment replicates) to genome-wide identification of differentially methylated loci information (chr1 to chr22, chrX, and chrY).

The second functional procedure is statistical analysis, genomic annotation, and functional information retrieval from the curated DNA methylation profiles, which outputs the differentially methylated CpG (DMC) or region (DMR) for pairwise samples and intercellular interrogation [19], together with the statistical property analysis of those output sources.

The last analysis procedure is the integration and visualization, which provides insightful clues for statistical comparison and further experiment validation; it aims to identify statistically significant DMRs with underlying biological functions of interest and annotate those DMRs with genetic transcript information, together with region-specific reference genome sequence information. Thus, such integrative comparison can shed light on the vital regulatory processes leading to carcinogenesis with a systematic approach.

In the following sections, we will demonstrate the major analysis procedures and corresponding statistical comparisons and integrative visualization on the curated DNA methylation data in RRBS format [17, 18], and we will identify DMR and classify the hyper- and hypo-DMR candidates [19] and implement function annotation for methylated CpG sites and regions.

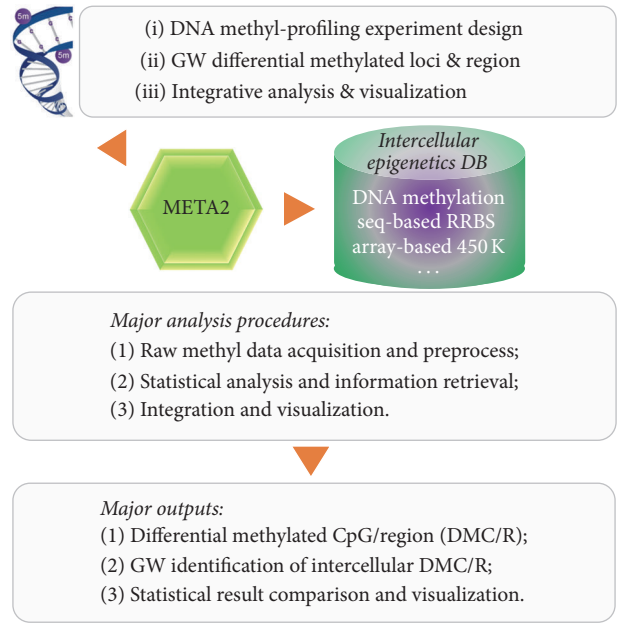

Figure 1: Schematic illustration for META2 structure and functions. META2 performs three major functional procedures, namely, DNA methylation raw data acquisition and preprocess (cell line curation and data format process), statistical analysis and information retrieval (CpG annotation, differential methylated CpG loci, and regions), and integration and results visualization (comparison and validation), together with the corresponding outputs as depicted on the bottom.

3. Comprehensive Analysis and Functional Annotation in META2

Here we propose the functions and analysis procedure in META2. As depicted in Figure 1, it mainly includes three major procedures as DNA methylation data source preprocess, information retrieval, and DNA methylation annotation. Thus, the below analysis results contain the following steps.

3.1. Statistics for Sequencing Read Coverage and Methylation Distribution. Firstly as for a high-throughput Next-Generation Sequencing (NGS) experiment, such as ChIP-seq or RRBS experiment, the necessary preprocess includes data quality check and preliminary statistical interrogation; thus biologists may gather the basic experiment quality information for following interrogation. Thus, we performed statistical calculation for the sequencing reads coverage counts (Cs and Ts) for the 1,135,337 CpG sites across the T-47D cell line.

Figure 2 illustrates the sequencing reads coverage information and DNA methylation distribution of RRBS data format for T-47D cell type. Figure 2 indicates that for both conditions' samples there exists the bimodal density pattern in the genome-wide methylation level with respect to the positive and negative strands, respectively.

And we also perform the genome-wide correlation analysis on the RRBS DNA methylation profile, and we find that there exists high correlation by pairwise comparison on control and treatment samples, with correlation coefficient from 0.94 to 0.96; see Figure 3(a).

Furthermore, we implement the region-specific analysis on those 1,135,337 CpG loci, and we find that genomic promoter and exon regions host more hypermethylated loci (≥25%) than hypomethylated loci (≤25% of methylation difference), which indicates that it is generally with hypermethylated status for most genes in T-47D cell. Together we also find that hypermethylated loci occur in CpG islands (59%) much more than hypomethylated loci (43%), which is basically consistent with the previous results; while CpG shores host 15% hypermethylated and 11% hypomethylated loci, respectively; see Figure 3(b).

Thus, based on the preprocess results, we perform the differential methylation analysis on those 1,135,337 CpG loci, and we get 3,651 statistically significant differentially methylated CpG loci (DMC), namely, absolute methylation difference ≥25% and its adjusted q-value ≤ 0.01. Those statistically significant DMCs provide meaningful clues for underlying genetic regulatory process when they are interrogated with further annotation and in silico deep analysis.

Thus, in the subsequent section, we will carry out differential methylated region (DMR) analysis on those identified

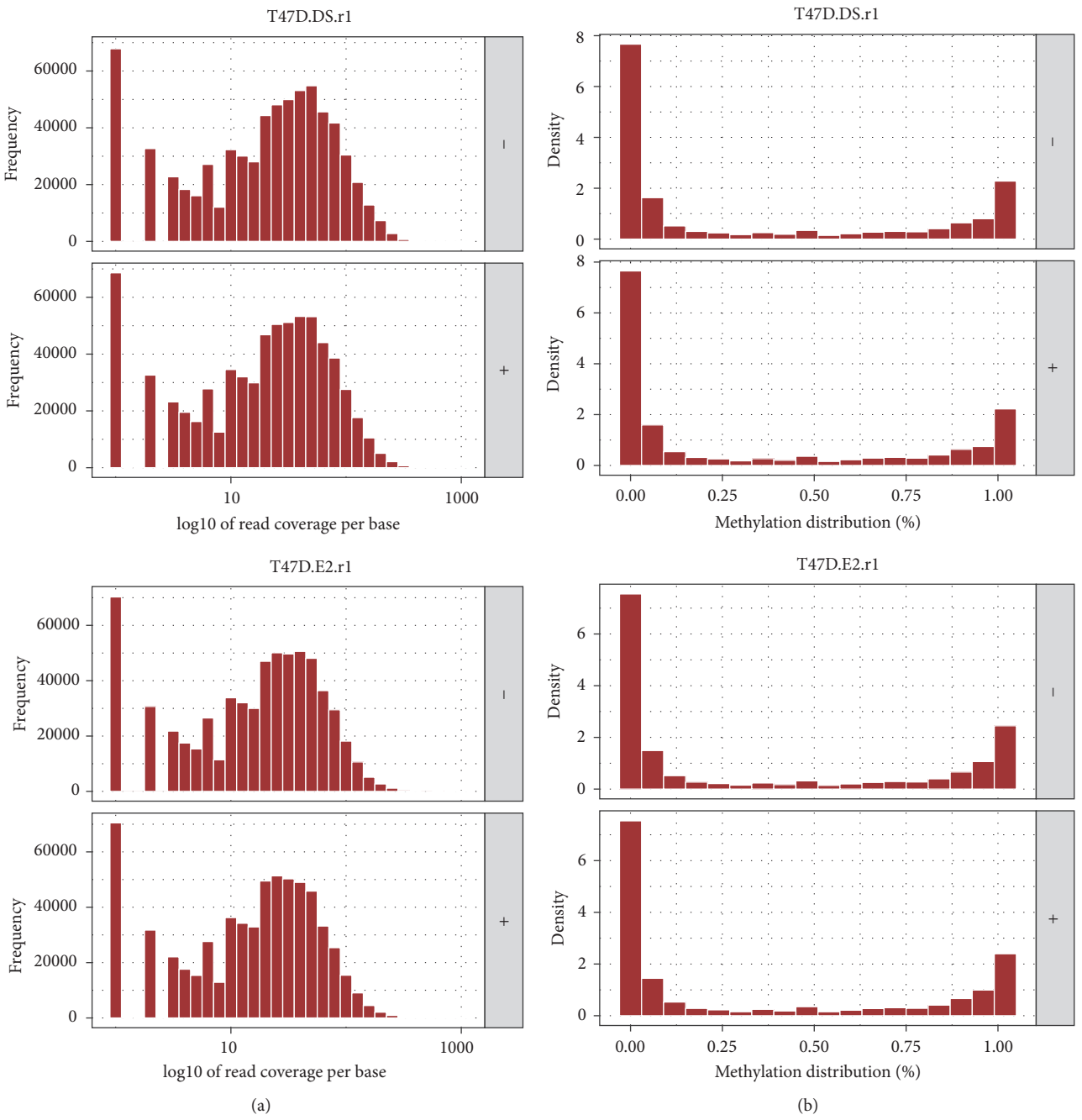

FIGURE 2: Schematic panel of statistical analysis on the raw RRBS data, that is, control (DS) versus treatment (E2) replicates with respect to positive and negative strands. (a) indicates the statistics for RRBS read coverage per base and (b) for the DNA methylation distribution for both control and treatment replicates.

DMCs, together with integrative analysis of genomic annotation.

3.2. Statistical Identification and Analysis of the Length-Specific DMRs. For consistence, we map genome-wide methylated CpG loci on each single chromosome (chr1 to chr22, chrX and chrY); see Figure 4, where each dot represents the differential CpG methylation level (in percentage, %) at the corresponding genomic position, and the line illustrates the general trend of differential methylation level across the whole chromosome.

We can see that chromosomes 1 and 2 host the longest differential methylation ranges, where the differentially methylated loci on chromosome 1 account for the most percentage (8.745%, 99,280 loci); the loci on chromosomes 17 and 19 also account for 6.506% (73,863 loci) and 6.959% (79,005 loci),

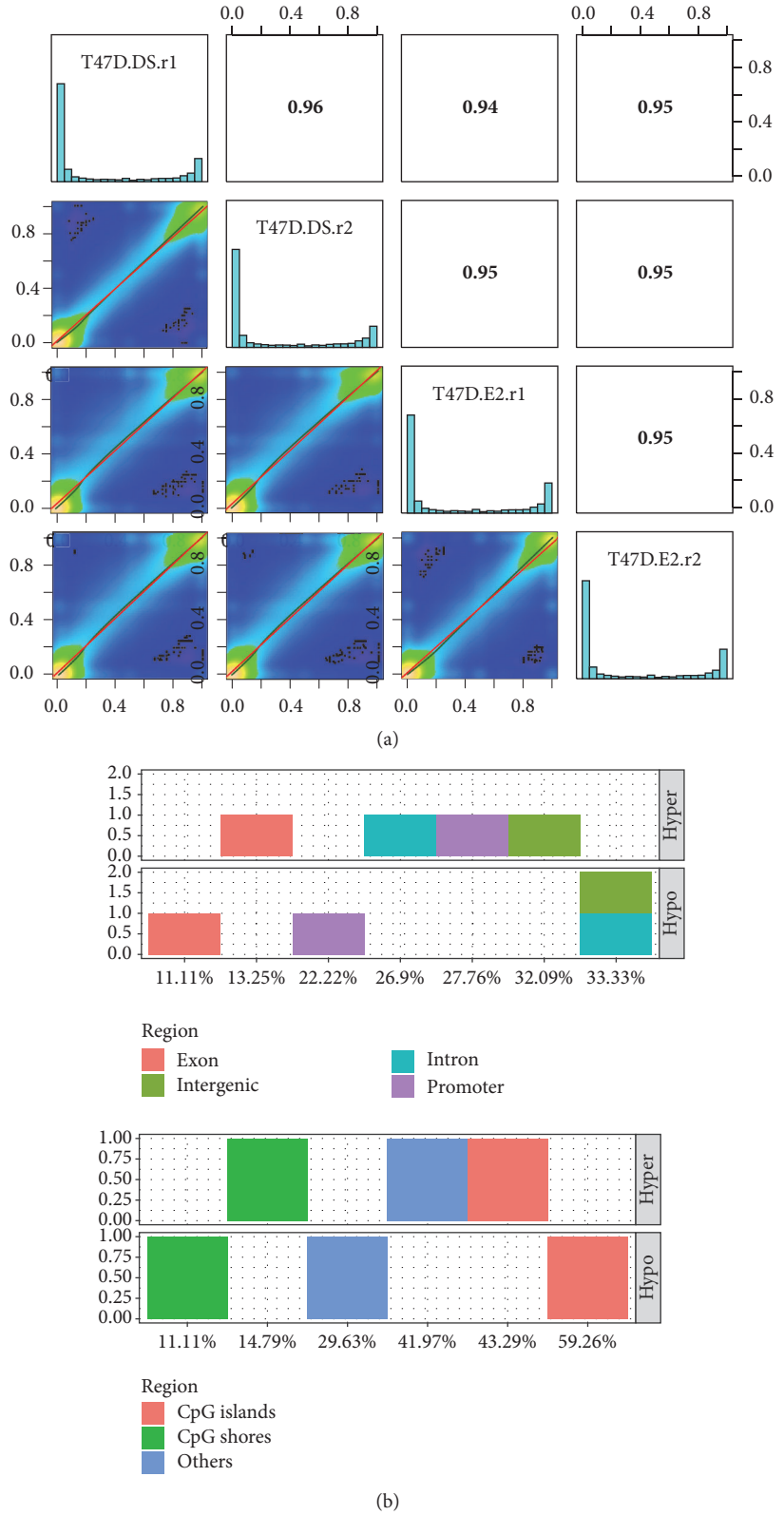

Figure 3: Schematic illustration of statistical correlation and methylation loci/region annotation analysis. (a) Correlation analysis for replicate methylation level (in percentage) from RRBS profiling technology; (b) genomic distribution for the differential methylated loci with respect to hyper- and hypomethylation status.

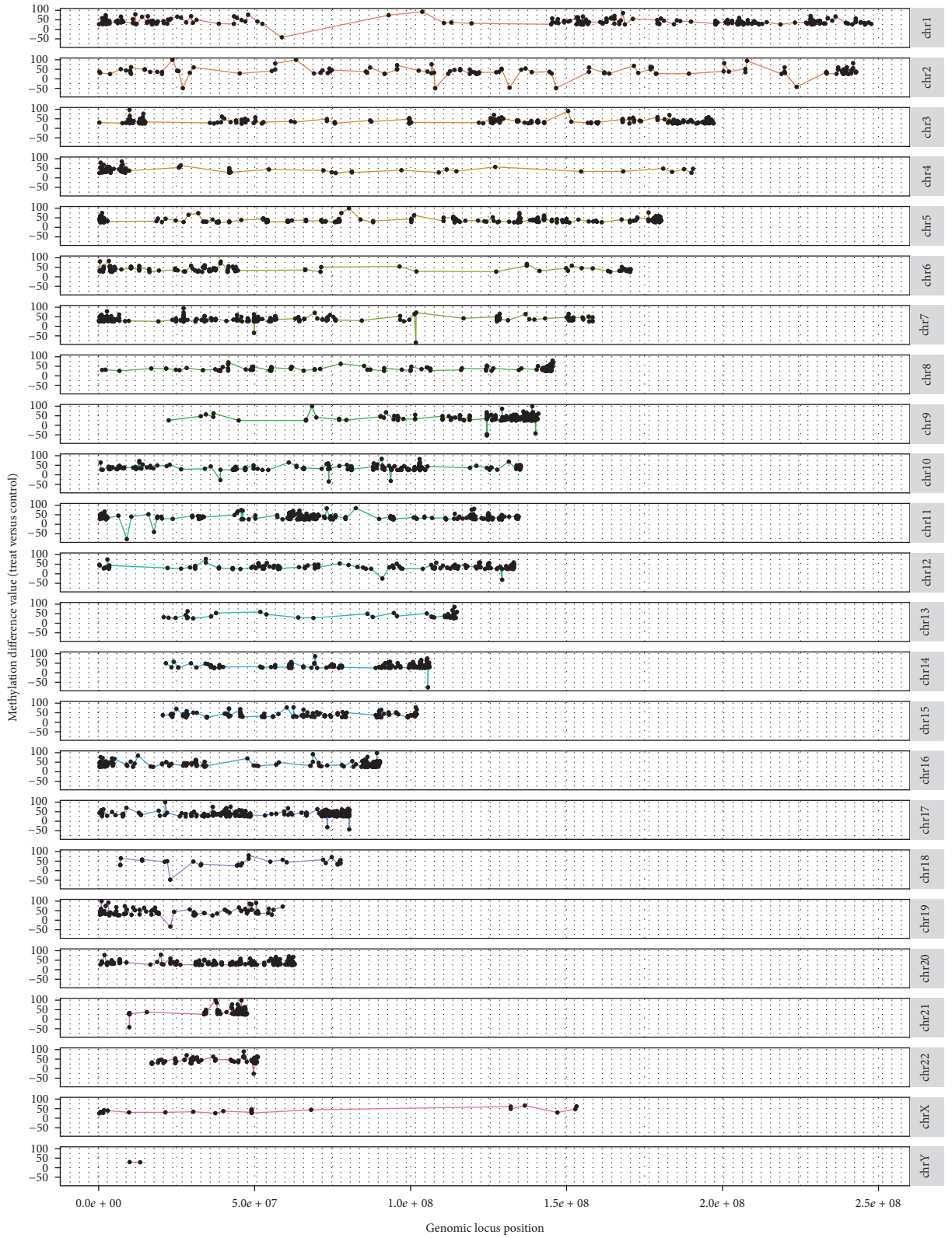

Figure 4: Genome-wide illustration for the identified differentially methylated CpG loci for T-47D cell type. Each black dot denotes the differentially methylated loci in base pair, and each curve depicts the general differential methylation trend for each chromosome.

ranking as the second and third, respectively, and those on chromosome 2 account for 6.351% (72,107 loci) of the total loci count (1,135,337).

Then we may question whether there exists any underlying biological function with those differential methylated loci, especially those with statistical significance and whether or not those loci have any clinical impact, and by which means? With those questions, META2 incorporates our self-compiled functions to interrogate the key points; firstly, META2 aims to uncover the differential methylation regions (DMRs) based on the statistically significant DMCs identified in the previous section.

We utilize a sliding window with a 10 bp length in scanning the whole genome to identify all DMR candidates. We predefine the DMR candidates that should cover more than two distinct but statistically significant DMCs to ensure the underlying biological meanings and also preset the generic DMR length up to 20,000 bp to interrogate the association between DMC count and DMR count with respect to DMR length.

Here we define two statistic indexes for measuring differential methylation level across multiple genomic loci and regions, namely, DMV.Sig for the highest differential methylation level of a significant DMC in a specific DMR and DMV.Avg for the averaged differential methylation level for all DMCs in a specific DMR. The index, DMV.Sig, aims to quantitatively identify the DMR candidates with significant methylation status across a specified range; DMV.Avg is for measuring the averaged methylation level within the DMR length under investigation.

Based on the change trends of DMV.Sig and DMV.Avg, we further utilize the information-theoretic measures, *Pearson* correlation and mutual information, for interrogating the region-specific methylation level; both of the measures intend to capture the statistical properties of dynamic variation in differential methylation profile.

Thus we calculate and illustrate the statistical association between DMR length and DMC/DMR count in Figure 5(a); Figure 5(b) depicts the statistical characteristics between mutual information and correlation analysis on DMV.Sig and DMV.Avg along with the DMR length.

In Figure 5(a), we find that along with DMR length up to 20,000 bp, DMC and DMR counts continue to increase (DMC count with a relatively sharper slope than DMR count), and for the region methylation indexes, DMV.Sig remains comparatively more stable (within the ranges 37.34% and 38.46%) than DMV.Avg, which decreases fleetly from 33.9% to 18.14%.

The analysis results above basically validate the hypothesis that DMR count statistically depends less than DMC count on the preset DMR length; meanwhile DMV.Avg depends more greatly than DMV.Sig on the preset DMR length, which indicates that the index DMV.Sig remains approximately the same for each DMR candidate regardless of the DMR length, while DMV.Avg decreases due to more and more low methylation loci covered by the subsequent DMR candidate.

Furthermore, from the results on the right panel (Figure 5), we find that both statistical curves undergo three critical transitions, that is, the shade zones A, B, and C. Zone A (DMR length at 1,500 bp) manifests the first transition at zero point for both mutual information and correlation coefficient, where DMV.Sig and DMV.Avg begin to take negative correlation; Zone B (DMR length at 5,000 bp) shows the second transition, where both mutual information and correlation for both indexes have evident inflections; Zone C (DMR length at 8,500 bp) indicates the third transition, where the negative correlation of both indexes begins to increase and their mutual information also rises up after a stretch of equilibrium. When the DMR length exceeds 12,500 bp, there is no apparent spinodal where both curves sustain the increase and decrease trends.

Based on those transition zone information, we can further annotate and decipher the underlying regulatory functions and biological meanings hereinafter.

3.3. Genomic Annotation and Identification of Genes Interacting with DMRs. Based on the three identified transition zones, we further implement genomic annotation and statistical analysis on the DMR candidates. For interrogating the inherent biology function, we emphasize the first transition zone; thus hereinafter we consider a specific class of DMRs with the maximum length less than 1,000 bp.

Figure 6 depicts those DMRs with relatively more differentially methylated loci; for illustration, we select nine typical DMRs from chromosomes 1 to 5.

From all the nine DMR distribution curves, we find that those DMRs mostly contain both hypermethylation and hypomethylation loci, while the former's count and differential methylation level are relatively more than the latter's count, which means those DMRs are generally with hypermethylation status. Subplot (g) is a good case in point, with nearly all loci being above the methylation level of 40%.

Meanwhile we further annotate the DMR candidate in subplot (b) with relatively more methylated loci than other DMR candidates. Figure 7 gives genomic annotation and analysis for the DMR in Figure 6(b), which is hosted in chromosome 1 and covers a 747 bp range from 197,743,880 to 197,744,626 bp. We acquire this DMR's methylation information and reference genome sequences from UCSC (hg19), together with protein-coding gene information.

From top to bottom panel, Figure 7 depicts this hyper-/hypomixture DMR genomic location in chromosome 1, as indicated in red line. The second panel depicts the DMR plot with 88 distinct methylation loci converged within the DMR, where those methylation loci constitute a hyper-/hypomixture methylation landscape directly impacting the underlying transcription regulatory processes for the targeted genes.

The third panel gives the reference genome sequence density within the exact DMR range; we can see that C/G content is comparatively higher than A/T in this DMR, which accords with the hypothesis that quite a few CpG sites cover the region. The bottom panel illustrated the five annotated transcripts for DENND1B at 1q31.3 (chr1:197,504,748–197,782,175), where the five transcripts generally maintain the hypermethylation status due to its range covering most hypermethylation loci with its differential methylation level up to 45%, together with a few hypomethylation loci around 10%.

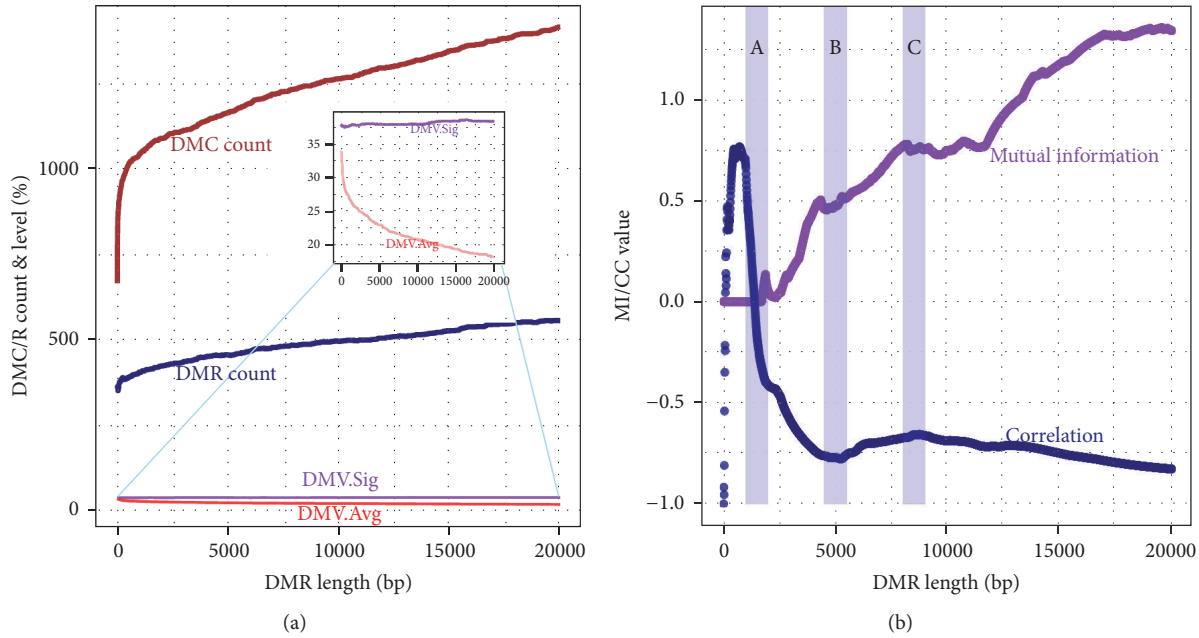

Figure 5: Statistical association analysis for DMC/R count, methylation level with respect to DMR length (a), and dynamic properties of mutual information (MI) and correlation coefficient (CC) with respect to DMR length (b).

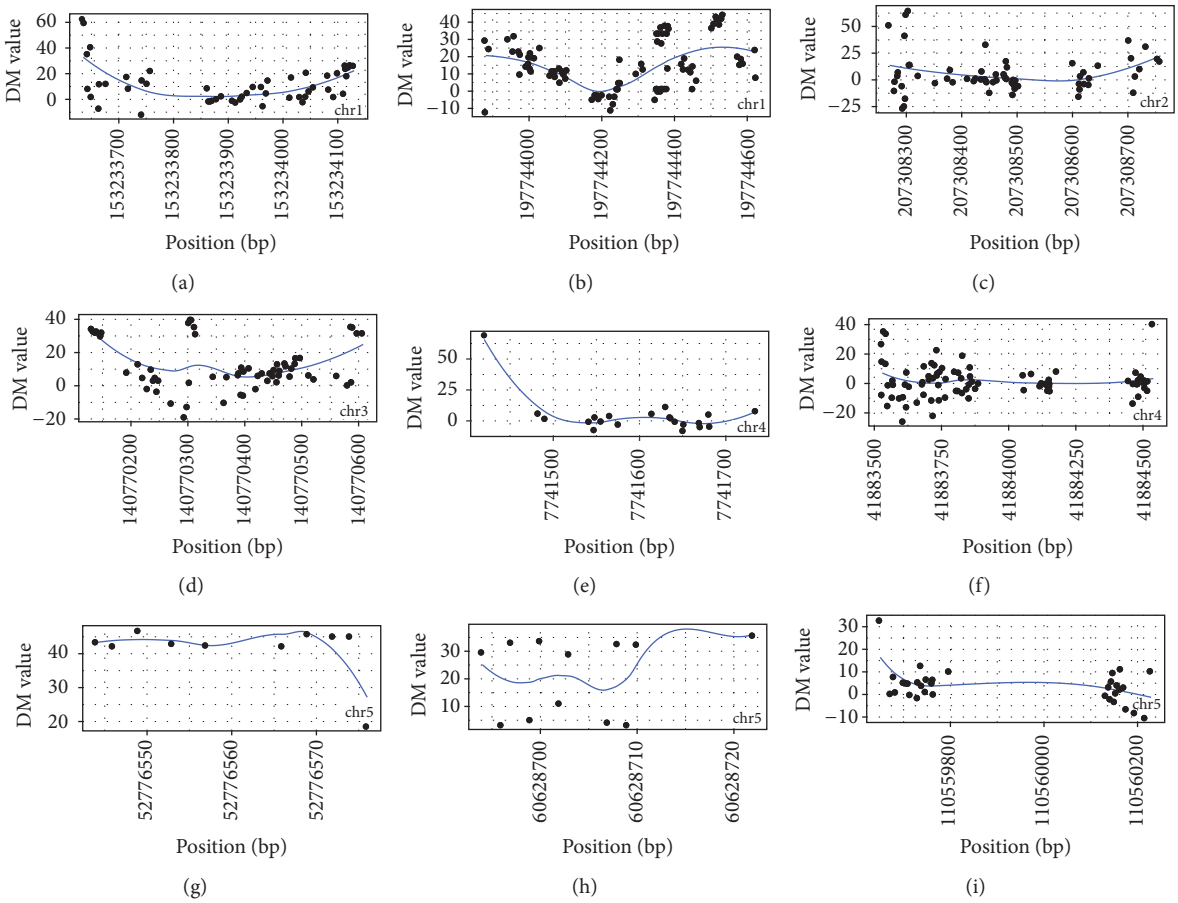

Figure 6: Schematic illustration for typical DMR candidates in chromosomes 1 to 5. The black dots denote the differential methylation value for each loci, and the blue lines represent the fitted DMR curves at each specific region. (a–i) subplots; (a) and (b) depict DMRs for chromosome 1, (c) for chromosome 2, (d) for chromosome 3, (e) and (f) for chromosome 4, and (g), (h), and (i) for chromosome 5.

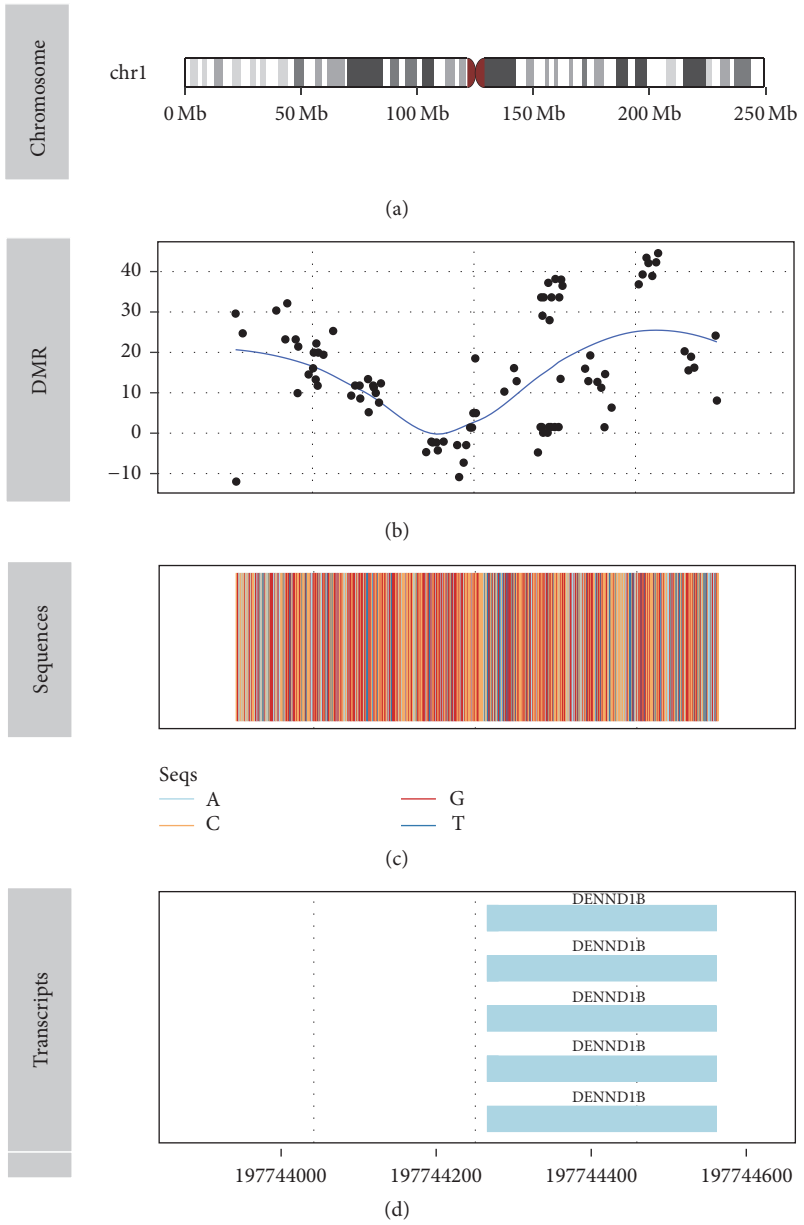

Figure 7: Schematic illustration of annotation and analysis for one typical DMR (chr1:197,743,880-197,744,626). From top to bottom, each panel depicts chromosome band, DMR and loci distribution, reference genome sequence (hg19), and annotated transcript information for DENND1B at 1q31.3.

4. Materials and Methods

4.1. Reduced Representation Bisulfite Sequencing (RRBS). Reduced representation bisulfite sequencing, or RRBS, is a large-scale random approach for analyzing and comparing genomic methylation patterns. BglII restriction fragments of 500–600 bp sized selected, together with adapters assembled, were further treated with bisulfite, PCR amplification, and clone and finally sequenced to target methylated CpG sites. From the converted and unconverted read counts at each CpG, the sample coverage and methylation level (in percentage) can be acquired [11–13].

4.2. Annotation for the Significant Differentially Methylated CpG Sites (DMC). Here we selected one cell line (T-47D, control versus treatment) as the benchmark cell line, and the annotation results are further filtered based on the lifted methylation difference threshold (at least 25% methylation difference for the paired groups).

4.3. Statistical Analysis for the Differentially Methylated Regions. We identified 16,277 DMR candidates from all the DMCs, with the adjusted q-value ≤ 0.01, CpG base methylation difference cutoff, 25, and DMR mean methylation difference cutoff, 20. Within those candidates, 8,936 entries present

hypermethylated and 7,341 hypomethylated status. With the lifted thresholds, namely, adjusted q-value ≤ 0.001 and differentially methylated CpG base count ≥ 5, we further detected 7,537 significant DMRs (Sig-DMRs), where 3,512 entries are significantly hypermethylated-DMRs (Sig-Hyper-DMRs) and 4,025 significantly hypomethylated-DMRs (Sig-Hypo-DMRs).

4.4. Tools Used in the Raw RRBS Curation and Statistical Analysis. Bowtie2 [20] was used to align sequencing reads; SAMtools [21] and BAMtools [22] were used to process the aligned sequencing reads, and methylKit [23] and META2 package were used to analyze the raw RRBS data; limma and DEseq were used in differential analysis of DNA methylation loci [24].

4.5. Generalized Mutual Information. Given two discrete random variables X and Y, the mutual information is defined as

$$\mathrm{MI}(X,Y) = H(X) + H(Y) - H(X,Y), \quad (1)$$

where $H(X)$ and $H(Y)$ are the entropy measures for X and Y and $H(X,Y)$ is the joint entropy between variables X and Y, respectively. The mutual information measure is adopted for association identification within the analysis section.

5. Conclusion

Here we present a developed toolkit, META2, for DNA methylation annotation and analysis, which aims to implement the intercellular analysis on differentially methylated loci and regions. META2 contains multiple versatile functions for annotating and analyzing DNA methylation, such as the profiling data by RRBS and other high-throughput technology.

By utilizing the toolkit on the real RRBS data from ENCODE, we performed statistical correlation and genomic loci/region annotation for all the identified differentially methylated CpGs, or DMC candidates; we further implemented statistical association analysis for DMC/R count and methylation level with respect to the preset DMR length and revealed the dynamic properties of mutual information and correlation coefficient with respect to DMR length; thus we detected three major transition zones, which provide statistical clues for further biological function investigation.

Our work provides a versatile and comprehensive analysis toolkit for epigenetic research and clinical study, especially for the genome-wide biomedical analysts, to interrogate and validate their hypothesis in an efficient and uniform way.

Further anticipated improvements including statistical annotation and analysis functions concerning cell or tissue-specific and pancancer analysis functions will be consolidated into the toolkit; thus it constitutes a versatile and evolving toolkit for biologists to easily adopt in their research.

Additional Points

Availability. RRBS profiling data for T47D is available at ENCODE; META2 and its corresponding test data and manual were deposited at https://github.com/gladex/META2.

Competing Interests

The author declares that there are no competing interests regarding the publication of this paper.

Acknowledgments

This work was supported by the Natural Science Foundation of Jiangsu, China (BE2016655 and BK20161196), the Fundamental Research Funds for China Central Universities (2016B08914), and Changzhou Science & Technology Program (CE20155050). 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- (OCC-) 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.

References

[1] The Cancer Genome Atlas Research Network, J. N. Weinstein, E. A. Collisson et al., "The cancer genome Atlas pan-cancer analysis project," *Nature Genetics*, vol. 45, pp. 1113–1120, 2013.

[2] V. N. Kristensen, O. C. Lingjærde, H. G. Russnes, H. K. M. Vollan, A. Frigessi, and A.-L. Børresen-Dale, "Principles and methods of integrative genomic analyses in cancer," *Nature Reviews Cancer*, vol. 14, no. 5, pp. 299–313, 2014.

[3] T. Witte, C. Plass, and C. Gerhauser, "Pan-cancer patterns of DNA methylation," *Genome Medicine*, vol. 6, no. 8, article 66, pp. 1–18, 2014.

[4] M. D. M. Leiserson, F. Vandin, H.-T. Wu et al., "Pan-cancer network analysis identifies combinations of rare somatic mutations across pathways and protein complexes," *Nature Genetics*, vol. 47, no. 2, pp. 106–114, 2015.

[5] The Cancer Genome Atlas Research Networ, "Comprehensive genomic characterization of squamous cell lung cancers," *Nature*, vol. 489, no. 7417, pp. 519–525, 2012.

[6] The Cancer Genome Atlas Research Network, "Comprehensive molecular portraits of human breast tumours," *Nature*, vol. 490, no. 7418, pp. 61–70, 2012.

[7] A. Meissner, T. S. Mikkelsen, H. Gu et al., "Genome-scale DNA methylation maps of pluripotent and differentiated cells," *Nature*, vol. 454, no. 7205, pp. 766–770, 2008.

[8] H. Heyn, F. Carmona Javier, A. Gomez et al., "DNA methylation profiling in breast cancer discordant identical twins identifies DOK7 as novel epigenetic biomarker," *Carcinogenesis*, vol. 34, no. 1, pp. 102–108, 2013.

[9] R. Lister, E. A. Mukamel, J. R. Nery et al., "Global epigenomic reconfiguration during mammalian brain development," *Science*, vol. 341, no. 6146, Article ID 1237905, 2013.

[10] J. T. Bell, A. K. Loomis, L. M. Butcher et al., "Differential methylation of the TRPA1 promoter in pain sensitivity," *Nature Communications*, vol. 5, article no. 2978, 2014.

[11] A. Meissner, A. Gnirke, G. W. Bell, B. Ramsahoye, E. S. Lander, and R. Jaenisch, "Reduced representation bisulfite sequencing for comparative high-resolution DNA methylation analysis," *Nucleic Acids Research*, vol. 33, no. 18, pp. 5868–5877, 2005.

[12] H. Guo, P. Zhu, L. Yan et al., "The DNA methylation landscape of human early embryos," *Nature*, vol. 511, no. 7511, pp. 606–610, 2014.

[13] A. Kundaje, W. Meuleman, J. Ernst et al., "Integrative analysis of 111 reference human epigenomes," *Nature*, vol. 518, no. 7539, pp. 317–330, 2015.

[14] J. Maksimovic, L. Gordon, and A. Oshlack, "SWAN: subset-quantile within array normalization for illumina infinium humanmethylation450 beadchips," *Genome Biology*, vol. 13, no. 6, 2012.

[15] M. Bibikova, B. Barnes, C. Tsan et al., "High density DNA methylation array with single CpG site resolution," *Genomics*, vol. 98, no. 4, pp. 288–295, 2011.

[16] R. Lister, M. Pelizzola, R. H. Dowen et al., "Human DNA methylomes at base resolution show widespread epigenomic differences," *Nature*, vol. 462, no. 7271, pp. 315–322, 2009.

[17] M. J. Ziller, H. Gu, F. Müller et al., "Charting a dynamic DNA methylation landscape of the human genome," *Nature*, vol. 500, no. 7463, pp. 477–481, 2013.

[18] A. Blattler, L. Yao, H. Witt et al., "Global loss of DNA methylation uncovers intronic enhancers in genes showing expression changes," *Genome Biology*, vol. 15, article 469, 2014.

[19] C. J. Kemp, J. M. Moore, R. Moser et al., "CTCF haploinsufficiency destabilizes DNA methylation and predisposes to cancer," *Cell Reports*, vol. 7, no. 4, pp. 1020–1029, 2014.

[20] B. Langmead and S. L. Salzberg, "Fast gapped-read alignment with Bowtie 2," *Nature Methods*, vol. 9, no. 4, pp. 357–359, 2012.

[21] H. Li, B. Handsaker, A. Wysoker et al., "The sequence alignment/map format and SAMtools," *Bioinformatics*, vol. 25, no. 16, pp. 2078–2079, 2009.

[22] D. W. Barnett, E. K. Garrison, A. R. Quinlan, M. P. Strömberg, and G. T. Marth, "Bamtools: A C++ API and toolkit for analyzing and managing BAM files," *Bioinformatics*, vol. 27, no. 12, Article ID btr174, pp. 1691–1692, 2011.

[23] A. Akalin, M. Kormaksson, S. Li et al., "methylKit: a comprehensive R package for the analysis of genome-wide DNA methylation profiles," *Genome Biology*, vol. 13, no. 10, article R87, 2012.

[24] S. Anders and W. Huber, "Differential expression analysis for sequence count data," *Genome Biology*, vol. 11, no. 10, article R106, 2010.