Original Research

Epigenomic analysis of 5-hydroxy-methylcytosine (5hmC) reveals novel DNA methylation markers for lung cancers

CrossMark

Zhihui Wang [a,b]; Mulong Du [c]; Qianyu Yuan [a,b]; Yichen Guo [b,d]; John N. Hutchinson [d]; Li Su [b]; Yinan Zheng [f]; Jun Wang [f]; Lorelei A. Mucci [e]; Xihong Lin [d]; Lifang Hou [f]; David C. Christiani [b,g,*]

[a] Harvard Graduate School of Arts and Sciences, Harvard University, Boston, MA, USA; [b] Department of Environmental Health, Harvard T.H. Chan School of Public Health, Harvard University, Boston, MA, USA; [c] Department of Biostatistics, Center for Global Health, School of Public Health, Nanjing Medical University, Nanjing, China; [d] Department of Biostatistics, Harvard T.H. Chan School of Public Health, Harvard University, Boston, MA, USA; [e] Department of Epidemiology, Harvard T.H. Chan School of Public Health, Harvard University, Boston, MA, USA; [f] Department of Preventive Medicine, Northwestern University Feinberg School of Medicine, Chicago, IL, USA; [g] Massachusetts General Hospital, Harvard Medical School, Boston, MA, USA

Abstract

Background: DNA methylation at the fifth position of cytosine (5mC) is a common epigenetic alteration affecting a range of cellular processes. In recent years, 5-hydroxymethylcytosine (5hmC), an oxidized form of 5mC, has risen broad interests as a potential biomarker for lung cancer diagnosis and survival.

Methods: We analyzed the epigenome-wide 5hmC profiles of paired lung tumor and adjacent normal tissues, using the TET-Assisted Bisulfite (TAB) array – Infinium MethylationEPIC BeadChip (EPIC) approach. The differentially methylated CpG sites were identified, and the biological relevance of 5hmC was assessed by differential methylation regions (DMR) analysis and gene set analysis (GSA).

Results: We observed global hypomethylation of 5hmC comparing tumor to normal tissues, and hypermethylated 5hmC were enriched in CpG islands and gene upstream. Comparison of 5hmC and 5modC (total methylation: 5mC + 5hmC) profiling showed low correlation, and only a small proportion of the significant 5hmC loci overlapped with the significant total methylation loci. GSA analysis suggested that 5hmC was mainly involved in biological processes such as cellular process, biological regulation, and metabolic process.

Conclusion: This is the first study to analyze the epigenome-wide 5hmC profiles among paired lung tumor and normal tissues. We observed global hypomethylation of 5hmC in lung cancers, and hypermethylated 5hmC enriched in CpG islands and gene upstream. We found that the genome-wide 5hmC levels do not correlate with the total methylation, and the GSA suggested different biological functions of 5hmC compared to 5modC. Overall, our results demonstrate the potential of 5hmC as a novel biomarker for lung cancer.

* Corresponding author at: Department of Environmental Health, Harvard T.H. Chan School of Public Health, 665 Huntington Avenue, Building I Room 1401, Boston, MA 02115, USA.
e-mail address: dchris@hsph.harvard.edu (D.C. Christiani).

Introduction

Lung cancer is the leading cancer-related cause of death in the world. There were an estimated 1.8 million lung cancer deaths globally [10,11] and 154 thousand deaths [45] in the United States in 2018 alone. The current five-year relative survival is about 19% [37], which is lower than most leading cancers and has improved only slightly over the past four decades. The five-year relative survival for localized tumors is 56%; however, only 16% of all lung cancer cases are diagnosed at an early stage and the five-year relative survival for distant tumors is only 5%. Therefore, early detection and intervention remain to be some of the most important strategies to improve long-term outcomes.

The rapid advancement of affordable `omics'-based technologies has allowed faster identification of putative biomarkers for improving lung cancer diagnosis and prognosis [48]. As one of the most important epigenetic mechanisms, DNA methylation at the fifth position of cytosine (5mC) plays a vital role in many cellular processes, such as embryonic development, gene expression, genomic imprinting, X-chromosome inactivation, and chromatin structure and stability [9,12,28,33,41]. DNA methylation abnormalities, such as global hypomethylation [13] and hypermethylation at promoters of tumor suppressor genes [18], are found in most tumors including lung cancer.

In recent years, 5-hydroxymethyl-cytosine (5hmC), an oxidized form of 5mC by Fe(II) and α-KG-dependent Ten-eleven translocation (TET) dioxygenases [24,47], has risen broad interests. It has been purported that 5hmC could act as an intermediate in DNA demethylation, either through active TET-assisted oxidation [20], or passive replication dependent loss [7,19]. Others also suggested that 5hmC could confer a distinct epigenetic function independent of DNA demethylation [7].

Early studies were made through global detection approaches, and global loss of 5hmC has been commonly observed across a spectrum of human tumors [15,22,27,30,51]. Mutations in IDH, SDH, and FH genes can control 5hmC genomic levels and trigger carcinogenesis by regulating TET activity [5,8,23,31].

Given the important epigenetic functions of 5hmC and its interplay with 5mC, it's essential to develop accurate and reliable methods to distinguish 5hmC from 5mC at single-base resolution, to better understand the biological mechanisms of 5hmC in cancer etiology and progression. However, the conventional bisulfite sequencing, which is the gold standard for methylation analysis, cannot distinguish 5hmC from 5mC. In the study, we analyzed the whole-genome 5hmC profiles of paired tumor and adjacent normal tissues from 8 patients with primary lung squamous cell carcinoma (LUSC), using a state-of-the-art technique of the TET-Assisted Bisulfite (TAB) array [52] with Infinium MethylationEPIC BeadChip (EPIC) [34]. We propose to use this as an exploratory study to assess the potential of using 5hmC as a novel epigenetic biomarker for lung cancer, and to lay the foundation for future clinical assay development and validation studies.

Material and methods

DNA samples

Genomic DNA was extracted for eight lung squamous cell carcinoma (LUSC) tissues and adjacent normal tissues. Tissues samples were obtained from lung cancer patients in the Harvard Lung Cancer Study (HLCS) who underwent surgical excision of tumors at Massachusetts General Hospital or Dana-Farber Cancer Institute. Manual microdissections of 5-u histopathologic sections have been conducted by one of our anatomic pathologists. Each specimen was evaluated for the amount and quality of tumor cells and histological classification was conducted by a pathologist using the WHO criteria. Tissue DNA was extracted from 50 mg of frozen tumor or normal tissue using the Maxwell 16 tissue DNA purification kit (Promega, USA). Tumor purity has been analyzed using LUMP analysis [4], which uses the mean normalized methylation level of 44 CpGs found to be unmethylated in blood cells.

5hmC and 5mC profiling

Illumina MethylationEPIC with bisulfite and TET-assisted bisulfite treatment

The 5mC and 5hmC profiling was completed at the Northwestern University Population Epigenetics Laboratory. A total of 1000 ng genomic DNA (gDNA) was split into two aliquots of 500 ng processed through either traditional bisulfite conversion (Bis) or TET-assisted bisulfite conversion (TAB) (WiseGene, Inc., Chicago, IL) [36]. As shown in Fig. 1, bisulfite treatment alone converted unmethylated cytosine into uracil that would be read as thymine (T) after PCR amplification, whereas both 5mC and 5hmC were resistant to conversion and would be read as cytosine (C). The TAB workflow involved protection of 5hmC with β-glucosyltransferase, followed by TET-assisted oxidation of unprotected 5mC to 5caC, which would be converted to uracil by bisulfite treatment and be read as thymine just as unmethylated cytosine. Therefore, Bis workflow give readout of total methylation (5mC + 5hmC), while TAB workflow gives direct readout of 5hmC. Converted DNA of both workflows was input into the Illumina MethylationEPIC BeadChip (Illumina, Inc., San Diego, CA) [40].

#### Whole genome bisulfite sequencing with Illumina Methyl-Seq

To validate our results and to establish the feasibility of using next generation sequencing (NGS) to profile 5hmC in lung tissues, the total methylation profiling was repeated among 5 pairs of tumor and adjacent normal tissues using the targeted bisulfite sequencing (Methyl-Seq). Briefly, 1000 ng of gDNA was sheared to yield fragment DNA between 100–300 base pair. These fragment DNA was further end-repaired, 3′-end adenylated, ligated, and hybridized; and captured DNA was bisulfite-converted using the Illumina TruSeq Methyl Capture EPIC Library Prep Kit (Illumina, Inc., San Diego, CA) [32]. Last, DNA was PCR-amplified and sequenced on Illumina Hi-Seq2500 System.

Statistical analysis

Illumina MethylationEPIC Bis/TAB data

The raw Tab-array and EPIC array data in IDAT format were read directly into R (version 3.2.4) and were processed using the Chip Analysis Methylation Pipeline (ChAMP) [35].

Quality control was conducted separately for the Bis and TAB workflow using the same criteria. We excluded unqualified probes, including: 1) probes with detection p value less than 0.01 in one or more samples (N_{Bis} = 1476, N_{TAB} = 84,125); 2) probes with a bead count < 3 in at least 5% of samples (N_{Bis} = 14,963, N_{TAB} = 23,966); 3) Non CpG probes (N_{Bis} = 2947, N_{TAB} = 2526); 4) SNP-related probes [53] (N_{Bis} = 96,512, N_{TAB} = 87,925); and 5) multi-hit probes [38] (N_{Bis} = 11, N_{TAB} = 10). Note that the number of probes removed due to low detection p value is much higher for 5hmC as the genome-wide 5hmC level is much lower than that of 5mC. Because the p value filter was applied first, the number of probes deleted for annotation purposes were different for 5hmC and 5-modC accordingly. After filtering, 750,009 and 667,366 probes were included for Bis and TAB workflow respectively. Normalization was performed to minimize unwanted variation within and between samples and we chose peak-based correction (PBC) [11], which has been shown as the most effective method when comparing with the whole genome bisulfite sequencing results [49].
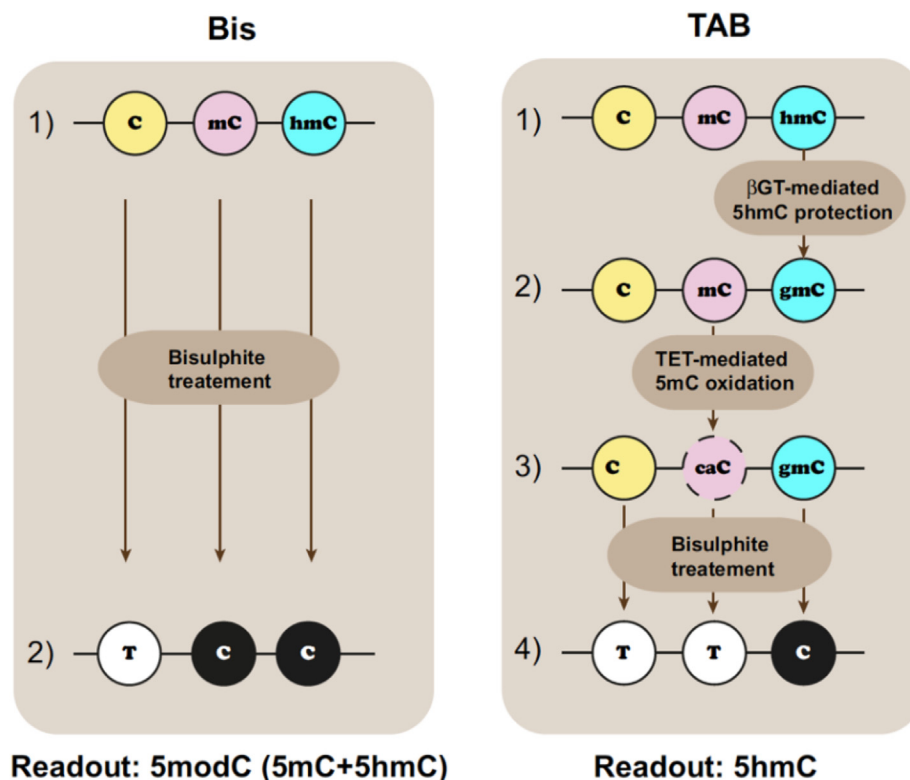
Fig. 1. The workflow of Bis and Tab conversion. Adapted from by Skvortsova et al. [44]. Copyright 2017 by BioMed Central.

For each CpG site, the methylated intensity (M) and unmethylated intensity (U) was transformed to $\beta = \frac{M}{M+U}$. The beta values with Bis conversion represent the proportion of total methylation (5modC = 5mC + 5hmC), whereas the beta values with TAB conversion represent the proportion of 5hmC methylation for each CpG site. The M values were calculated as $M = log_2(\frac{M}{U})$ and were used for regression analyses. To detect the differential methylation patterns between tumor and normal tissues, differential methylation probes (DMP) analysis were conducted using the paired t-test with Limma [42]. The DMPs were grouped into clusters/regions and differential methylated regions (DMRs) were identified using the Bumphunter algorithm [21]. The functional relevance of 5hmC/5modC-containing genes was explored using Gene Set Analysis (GSA) [39] with 15,996 Gene Ontology Biological Process (GOBP) pathways [6], while accounting for the number of probes per gene [14]. False discovery rate (FDR) of <0.05 was used in all analyses to account for multiple comparisons.

Whole genome bisulfite sequencing data

Whole genome bisulfite sequencing data was analyzed in collaboration with Harvard Chan Bioinformatics Core. FASTQC (Version 0.11.8) [3] was used for quality control to examine sequence counts, sequence quality histograms, per sequence quality scores, per base sequence GC content, per base N content, sequence length distribution, sequence duplication levels, overrepresented sequences, and adapter content. We observed very low proportion of cytosine bases in R1s (forward reads or read-1 s of the read pairs), suggesting good bisulfite conversion. Reads were trimmed using Trim Galore (Version 0.6.0) [25] after assessment by M-bias plot [29] to remove low quality sequences, adapter sequences, and biased sequences. FASTQC was rerun after trimming to confirm efficacy. Bismark (version 0.21.0) [26] was used to align the reads to the Ensembl GRCh37 Homo sapiens assembly [2], and duplicates reads removed using Picard Tools (http://broadinstitute.github.io/picard/). Bismark also

summarized the cytosine counts at cytosines followed by guanines (CpGs), cytosine followed by non-guanines followed by guanines (CHGs), and cytosines followed by at least two non-guanines (CHHs). We observed very low proportion (<0.2%) of CHG and CHH methylation, suggesting good bisulfite conversion.

CpG methylation percentages and read coverage of CpG generated from Bismark was assessed for adequate coverage (>10) and expected methylation percentage distributions in R using methylKit (version 1.8.1) [1]. The Bismark generated data was then imported into R using the bsseq Bioconductor package [16] and annotated with Ensembl GRCh37 Homo sapiens assembly using annotatr (version 1.8.0). Differential methylation analysis for the sequencing data was conducted using a general linear model accounting for tumor-normal pairing in DSS (version 2.30.1) [50], a new dispersion shrinkage method that estimates the dispersion parameter from beta-binomial distribution. Again, test statistics were calculated at each differentially methylated loci (DMLs) and significant DMLs were aggregated into DMRs.

Results

Study subjects and sample characteristics

Tissue samples were obtained from 8 Caucasian males with early stage (stage 1 and 2) lung squamous cell carcinoma. The mean age at diagnosis is 69.2 years old (range = 63.1–76.9 years).

Using the LUMP method for tumor purity, the average tumor purity is around 40%, with the average normal tissue purity to be around 39%.

*Genomic alterations of 5hmC among lung tissues*

Using the TAB – EPIC and Bis – EPIC array, we profiled the 5hmC and 5modC levels in 8 pairs of tumor and adjacent normal samples

(Supplementary Fig. 1). While the distribution of total methylation demonstrated a typical two-mode distribution with peaks at both low and high β levels (Supplementary Fig. 1B), the distribution of 5hmC showed a unimodal distribution with the β values of most probes to be less than 0.1 (Supplementary Fig. 1A).

We further examined the genomic 5hmC and 5modC alterations by CpG island (CGI) status as well as genomic features. With regards to CGI status, we observed similar patterns for both 5hmC and total methylation: there was higher likelihood of hypomethylation at opensea and shelf regions as well as hypermethylation at shore and CGI regions (Supplementary Fig. 1C and D). For genomic features, 5hmC was hypermethylated at gene promoter regions (TSS1500 and TSS200), 5′UTR, and the 1st exon, but hypomethylated at gene bodies (Supplementary Fig. 1E).

Locus-specific alterations of 5hmC in lung cancers

We identified 25,611 differentially methylated loci for 5hmC at FDR < 0.05, among them 21,251 loci (83.0%) were hypomethylated; as compared to 5modC, there were 34,975 differentially methylated loci at FDR < 0.05. among them 30,616 loci (87.5%) were hypomethylated. Further examination of the distribution pattern of significant 5hmC loci by CGI status revealed that significant hypomethylated 5hmC loci were overrepresented in opensea (OR = 2.22, 95% CI = 2.15–2.28) whereas hypermethylated 5hmC loci were enriched in CGI (OR = 4.26, 95% CI = 4.01–4.52) (Fig. 2A). With regards to genomic features, hypomethylated 5hmC loci were more likely to distribute at gene bodies (OR = 1.54, 95% CI = 1.50–1.59), whereas hypermethylated 5hmC loci were more
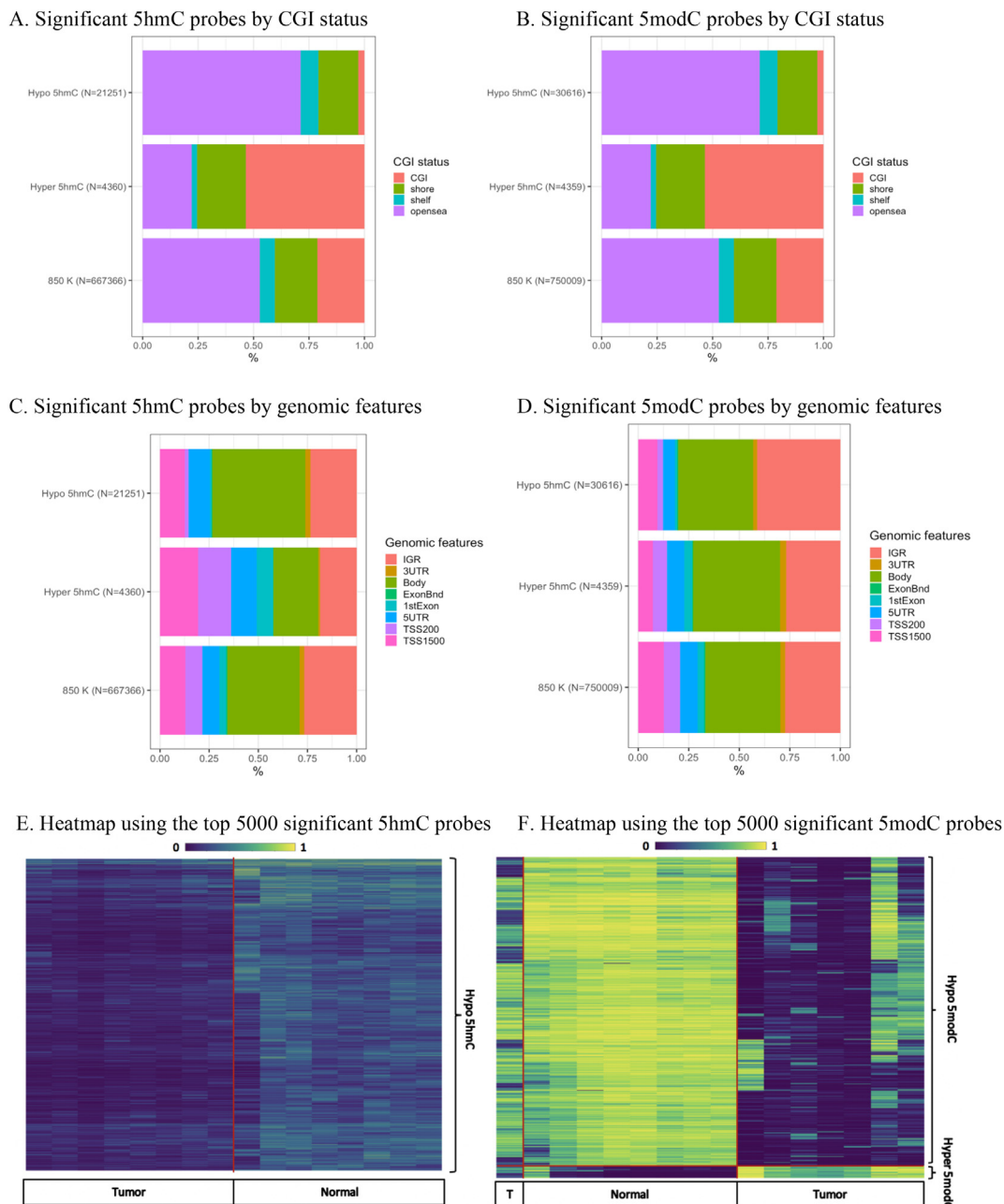
Fig. 2. Locus-specific alterations of 5hmC and total methylation in lung cancers. A. Significant 5hmC probes by CGI status. B. Significant 5modC probes by CGI status. C. Significant 5hmC probes by genomic features. D. Significant 5modC probes by genomic features. E. Heatmap using the top 5000 significant 5hmC probes. F. Heatmap using the top 5000 significant 5modC probes.
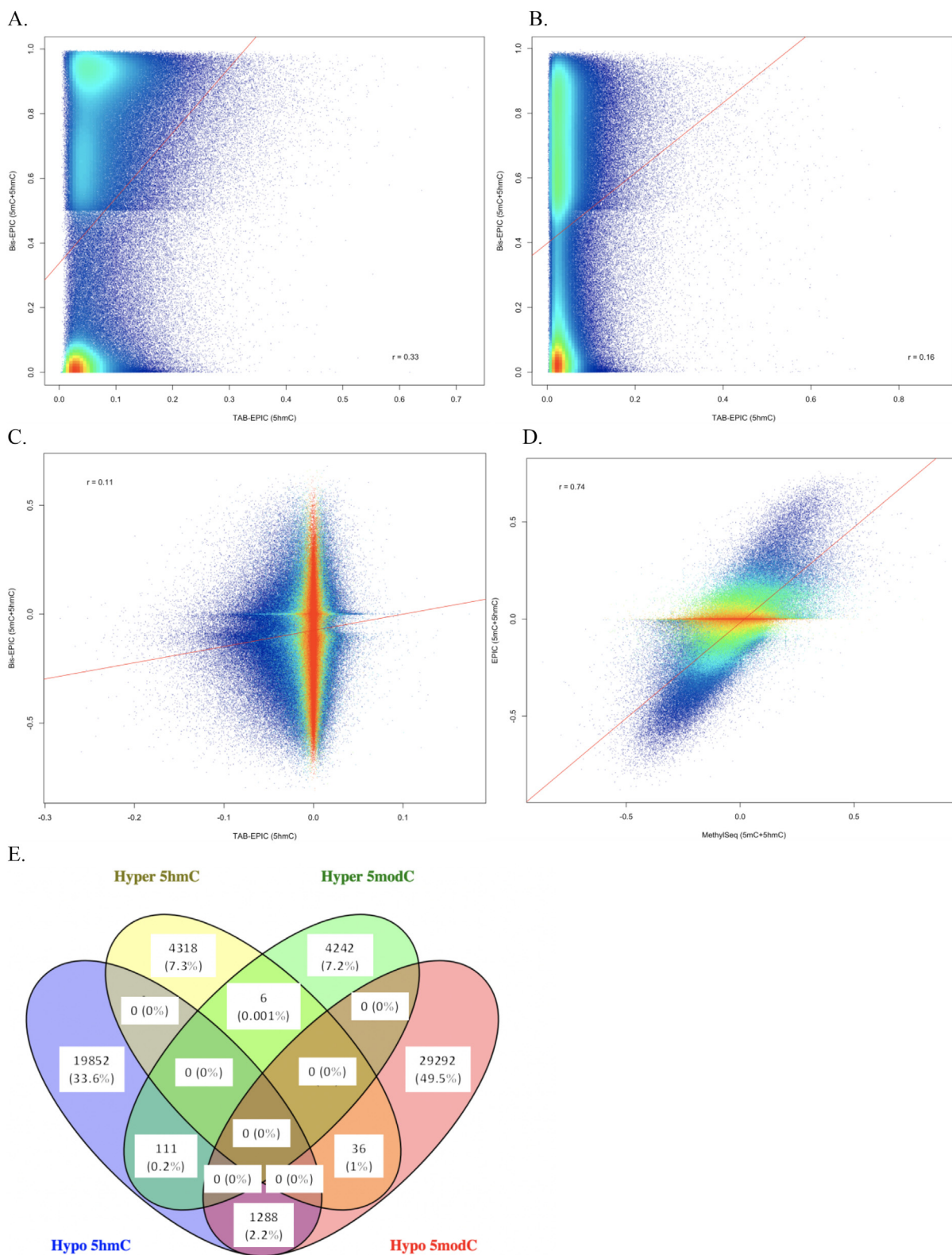
Fig. 3. Comparison of 5modC and 5hmC profiling. A. Scatter plot of βs between TAB-EPIC (5hmC) and Bis-EPIC (5modC) at CpG sites with significant differential hydroxymethylation. B. Scatter plot of βs between TAB-EPIC (5hmC) and Bis-EPIC (5modC) at CpG sites with significant differential total methylation. C. Scatter plot of average methylation difference between TAB-EPIC (5hmC) and Bis-EPIC (5modC). D. Scatter plot of average methylation difference between Bis-EPIC (5modC) and MethylSeq (5modC). E. Venn diagram comparing the CpG sites with significant differential total methylation and differential hydroxymethylation.

frequent at gene promoter regions (OR = 2.07, 95% CI = 1.95–2.21), 5′UTR (OR = 1.55, 95% CI = 1.42–1.70), and the 1st exon (OR = 2.57, 95% CI = 2.31–2.87) (Fig. 2C). Using the top 5000 significant probes for 5hmC, we could clearly distinguish normal and tissue samples with hierarchical clustering (Fig. 2E), and the discriminatory power of 5hmC model appeared to be superior than the total methylation model (Fig. 2F).

Comparison of 5hmC and 5modC profiling

We further compared the 5hmC and 5modC profiling measured by different methods. As 5hmC was considered as an intermediate of DNA demethylation, we examined the correlation between 5hmC and 5modC across CpG sites with significant differential hydroxymethylation (N = 25,611) (Fig. 3A) and across CpG sites with significant differential total methylation (N = 34,975) (Fig. 3B). The 5hmC levels at the significant CpG sites showed poorly correlation with the total methylation, suggesting potential distinct epigenetic function of 5hmC that is independent of 5mC. The average differences of 5hmC between tumor and normal tissues were also poorly correlated (r = 0.11) with the average difference of 5modC across genome (N = 656,768), though the average differences in 5modC measured by different methods (EPIC and MethylSeq) showed fairly high concordance (r = 0.74). The Venn diagram comparing the significant 5hmC loci (N = 25,611) and the significant 5modC loci (N = 34,975) found only 1441 (2.4%) overlapping CpG sites for both biomarkers, and among them 1288 loci (89.4%) were hypomethylated for both 5hmC and 5modC (Fig. 3E).

Exploring biological relevance of 5hmC loci

We aggregated probes into regions and identified 26 regions with differentially methylated 5hmC at FDR < 0.01, and 875 regions with differentially methylated 5modC at FDR < 0.05. Supplementary Table 1 shows the top 10 differential methylation regions (DMRs) for 5hmC and 5-modC. Though globally hypomethylated, 5hmC were hypermethylated at the upstream (gene promoter regions and 5′UTR) of top genes such as DYNLRB1, ALS2, CCNYL1, ATAD1, and STX16.

We conducted gene set analysis (GSA) to understand the biological processes that the differentially methylated probes/regions are involved in. In total, 430 and 189 GOBP pathways were significant for 5hmC and 5modC at FDR < 0.05 level, respectively. The top 10 significant pathways for 5hmC and 5modC are shown in Supplementary Table 2. We further explored the biological relevance of 5hmC by categorizing the GOBP pathways into 17 top-level categories (cellular process, biological regulation, metabolic process, multicellular organismal process, developmental process, response to stimulus, localization, cellular component organization or biogenesis, multi-organism process, signaling, immune system process, reproduction, locomotion, cell population proliferation, growth, biological adhesion, and behavior). For each category, we calculated the expected number of significant pathways and compared with the observed number of significant pathways in each category (Fig. 4). We found that the differentially methylated 5hmC genes were more likely to congregate at pathways involved in cellular process, biological regulation, and metabolic process, whereas the differentially methylated 5modC genes were more frequent among pathways related to localization, signaling, and multicellular organismal process.

Discussion

Though the incidence of lung cancer has decreased by 3% per year in men and 1.5% per year in women from 2011 to 2015 in the US [46], it remains to be the most common type of cancer and the leading cause of
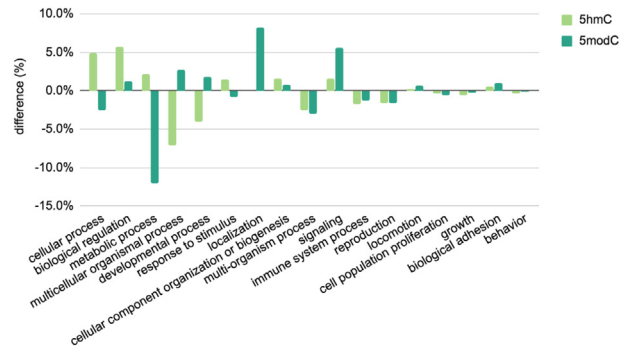

Fig. 4. Difference observed and expected number of significant pathways in major Gene Ontology Biological Process.

cancer-related death globally [10]. What's more, the current five-year relative survival of 19% is lower than most common cancers and has only improved slightly from 12% over the past three decades [37]. Therefore, sensitive and specific biomarkers for lung cancer are in urgent needs to improve diagnosis and long-term survival.

Due to the high tissue specificity and availability of high-throughput method for quantification, DNA methylation has been considered as a promising cancer biomarker and has been studied extensively in most cancers [17]. However, most studies on DNA methylation and cancer are based on bisulfite conversion methods, which cannot distinguish between 5hmC and 5mC. In this study, we applied a start-of-the-art technique of TET-Assisted Bisulfite (TAB) array with Illumina MethylationEPIC (EPIC) BeadChip, and quantified the novel 5hmC marker at ∼850,000 CpG sites across human genome.

In this study, we observed a global depletion of 5hmC with most CpG sites hypomethylated in tumors, which agrees with previous studies [15,22,27,30,51]. Notably, we found hypermethylation of 5hmC among tumors that were enriched in gene upstream (e.g. promoters), and previous *in vitro* studies reported that the presence of 5hmC at the promoter strongly suppressed transcription [43].

Though historically considered as an intermediate of DNA demethylation, the locus-specific 5hmC levels at differentially methylated CpG sites do not correlate with the 5modC levels, indicating that 5hmC may not simply be a demethylation intermediate. Moreover, comparison of the significant 5hmC loci with the significant 5modC loci found very few overlaps, and the differentially methylated 5hmC and 5modC genes were enriched in different biological systems, suggesting the potential of 5hmC as a new, independent biomarker for lung cancer diagnosis.

Conclusion

In this exploratory study, we analyzed the epigenome-wide 5hmC profiles among paired tumor and normal tissues using the TAB array – EPIC approach. We observed global hypomethylation of 5hmC in lung cancers, and hypermethylated 5hmC were enriched in CpG islands and gene upstream. We found that the genome-wide 5hmC levels do not correlate with total methylation, and gene set analysis revealed different biological functions of 5hmC. Overall, our results indicated the potential of 5hmC as a novel biomarker for lung cancer, and further work is needed to validate differentially methylated loci for cancer diagnosis and prognosis and to explore the functional role of 5hmC in lung cancer etiology.

Conflict of interest

The authors declare no conflicts of interest.

Acknowledgements

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.neo.2020.01.001.

References

1. Akalin A, Kormaksson M, Li S, Garrett-Bakelman FE, Figueroa ME, Melnick CE, Mason CE. methylKit: a comprehensive R package for the analysis of genome-wide DNA methylation profiles. *Genome Biol* 2012;**13**(10):R87. https://doi.org/10.1186/gb-2012-13-10-r87.

2. Aken BL, Ayling S, Barrell D, Clarke L, Curwen V, Fairley S . . . Searle SM. The Ensembl gene annotation system. *Database (Oxford)* 2016;**2016**. https://doi.org/10.1093/database/baw093.

3. Andrews S. FastQC: a quality control tool for high throughput sequence data; 2010.

4. Aran D, Sirota M, Butte AJ. Systematic pan-cancer analysis of tumour purity. *Nat Commun* 2015;**6**:8971.

5. Arvinden VR, Rao DM, Kuha A, Rajkumar T, Mani S. Regulation and functional significance of 5-hydroxymethylcytosine in cancer. *Epigenomes* 2017;**1**(3):19.

6. Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM . . . Eppig JT. Gene ontology: tool for the unification of biology. *Nat Genet* 2000;**25**(1):25.

7. Bachman M, Uribe-Lewis S, Yang X, Williams M, Murrell A, Balasubramanian S. 5-Hydroxymethylcytosine is a predominantly stable DNA modification. *Nat Chem* 2014;**6**(12):1049.

8. Bardella C, Pollard PJ, Tomlinson I. SDH mutations in cancer. *Biochim Biophys Acta* 2011;**1807**(11):1432–43. https://doi.org/10.1016/j.bbabio.2011.07.003.

9. Baylin SB, Esteller M, Rountree MR, Bachman KE, Schuebel K, Herman JG. Aberrant patterns of DNA methylation, chromatin formation and gene expression in cancer. *Hum Mol Genet* 2001;**10**(7):687–92.

10. Bray F, Ferlay J, Soerjomataram I, Siegel RL, Torre LA, Jemal A. Global cancer statistics 2018: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *CA Cancer J Clin* 2018;**68**(6):394–424. https://doi.org/10.3322/caac.21492.

11. Dedeurwaerder S, Defrance M, Calonne E, Denis H, Sotiriou C, Fuks F. Evaluation of the infinium methylation 450K technology. *Epigenomics* 2011;**3**(6):771–84. https://doi.org/10.2217/epi.11.105.

12. Eden A, Gaudet F, Waghmare A, Jaenisch R. Chromosomal instability and tumors promoted by DNA hypomethylation. *Science* 2003;**300**(5618):455. https://doi.org/10.1126/science.1083557.

13. Ehrlich M. DNA methylation in cancer: too much, but also too little. *Oncogene* 2002;**21**(35):5400.

14. Geeleher P, Hartnett L, Egan LJ, Golden A, Raja Ali RA, Seoighe C. Gene-set analysis is severely biased when applied to genome-wide methylation data. *Bioinformatics* 2013;**29**(15):1851–7.

15. Haffner MC, Chaux A, Meeker AK, Esopi DM, Gerber J, Pellakuru LG . . . Nelson WG. Global 5-hydroxymethylcytosine content is significantly reduced in tissue stem/progenitor cell compartments and in human cancers. *Oncotarget* 2011;**2**(8):627.

16. Hansen KD, Langmead B, Irizarry RA. BSmooth: from whole genome bisulfite sequencing reads to differentially methylated regions. *Genome Biol* 2012;**13**(10):R83. https://doi.org/10.1186/gb-2012-13-10-r83.

17. Hao X, Luo H, Krawczyk M, Wei W, Wang W, Wang J . . . Zhang K. DNA methylation markers for diagnosis and prognosis of common cancers. *Proc Natl Acad Sci U S A* 2017;**114**(28):7414–9. https://doi.org/10.1073/pnas.1703577114.

18. Herman JG, Baylin SB. Promoter-region hypermethylation and gene silencing in human cancer. *Curr Top Microbiol Immunol* 2000;**249**:35–54.

19. Inoue A, Zhang Y. Replication-dependent loss of 5-hydroxymethylcytosine in mouse preimplantation embryos. *Science* 2011;**334**(6053):194. https://doi.org/10.1126/science.1212483.

20. Ito S, Shen L, Dai Q, Wu SC, Collins LB, Swenberg JA . . . Zhang Y. Tet proteins can convert 5-methylcytosine to 5-formylcytosine and 5-carboxylcytosine. *Science* 2011;**333**(6047):1300–3.

21. Jaffe AE, Murakami P, Lee H, Leek JT, Fallin MD, Feinberg AP, Irizarry RA. Bump hunting to identify differentially methylated regions in epigenetic epidemiology studies. *Int J Epidemiol* 2012;**41**(1):200–9. https://doi.org/10.1093/ije/dyr238.

22. Jin S-G, Jiang Y, Qiu R, Rauch TA, Wang Y, Schackert G . . . Pfeifer GP. 5-Hydroxymethylcytosine is strongly depleted in human cancers but its levels do not correlate with IDH1 mutations. *Cancer Res* 2011;**71**(24):7360–5.

23. King A, Selak MA, Gottlieb E. Succinate dehydrogenase and fumarate hydratase: linking mitochondrial dysfunction and cancer. *Oncogene* 2006;**25**(34):4675–82. https://doi.org/10.1038/sj.onc.1209594.

24. Kriaucionis S, Heintz N. The nuclear DNA base 5-hydroxymethylcytosine is present in Purkinje neurons and the brain. *Science* 2009;**324**(5929):929–30.

25. Krueger F. Trim galore. A wrapper tool around Cutadapt and FastQC to consistently apply quality and adapter trimming to FastQ files; 2015.

26. Krueger F, Andrews SR. Bismark: a flexible aligner and methylation caller for Bisulfite-Seq applications. *Bioinformatics* 2011;**27**(11):1571–2.

27. Kudo Y, Tateishi K, Yamamoto K, Yamamoto S, Asaoka Y, Ijichi H . . . Koike K. Loss of 5-hydroxymethylcytosine is accompanied with malignant cellular transformation. *Cancer Sci* 2012;**103**(4):670–6.

28. Li E, Beard C, Jaenisch R. Role for DNA methylation in genomic imprinting. *Nature* 1993;**366**(6453):362–5. https://doi.org/10.1038/366362a0.

29. Lin X, Sun D, Rodriguez B, Zhao Q, Sun H, Zhang Y, Li W. BSeQC: quality control of bisulfite sequencing experiments. *Bioinformatics* 2013;**29**(24):3227–9. https://doi.org/10.1093/bioinformatics/btt548.

30. Liu C, Liu L, Chen X, Shen J, Shan J, Xu Y . . . Bie P. Decrease of 5-hydroxymethylcytosine is associated with progression of hepatocellular carcinoma through downregulation of TET1. *PLoS One* 2013;**8**(5) e62828.

31. Losman J-A, Kaelin WG. What a difference a hydroxyl makes: mutant IDH, (R)-2-hydroxyglutarate, and cancer. *Genes Dev* 2013;**27**(8):836–52.

32. Masser DR, Stanford DR, Freeman WM. Targeted DNA methylation analysis by next-generation sequencing. *J Vis Exp* 2015;**96**. https://doi.org/10.3791/52488.

33. Mohandas T, Sparkes RS, Shapiro LJ. Reactivation of an inactive human X chromosome: evidence for X inactivation by DNA methylation. *Science* 1981;**211**(4480):393–6.

34. Moran S, Arribas C, Esteller M. Validation of a DNA methylation microarray for 850,000 CpG sites of the human genome enriched in enhancer sequences. *Epigenomics* 2016;**8**(3):389–99. https://doi.org/10.2217/epi.15.114.

35. Morris TJ, Butcher LM, Feber A, Teschendorff AE, Chakravarthy AR, Wojdacz TK, Beck S. ChAMP: 450k chip analysis methylation pipeline. *Bioinformatics* 2014;**30**(3):428–30. https://doi.org/10.1093/bioinformatics/btt684.

36. Nazor KL, Boland MJ, Bibikova M, Klotzle B, Yu M, Glenn-Pratola VL . . . Fan JB. Application of a low cost array-based technique – TAB-Array – for quantifying and mapping both 5mC and 5hmC at single base resolution in human pluripotent stem cells. *Genomics* 2014;**104**(5):358–67. https://doi.org/10.1016/j.ygeno.2014.08.014.

37. Noone A, Howlader N, Krapcho M, Miller D, Brest A, Yu M . . . Lewis D. *SEER cancer statistics review, 1975–2015.* Bethesda, MD: National Cancer Institute; 2018.

38. Nordlund J, Bäcklin CL, Wahlberg P, Busche S, Berglund EC, Eloranta M-L . . . Harila-Saari A. Genome-wide signatures of differential DNA methylation in pediatric acute lymphoblastic leukemia. *Genome Biol* 2013;**14**(9) r105.

39. Phipson B, Maksimovic J, Oshlack A. missMethyl: an R package for analyzing data from Illumina's HumanMethylation450 platform. *Bioinformatics* 2015;**32**(2):286–8.

40. Pidsley R, Zotenko E, Peters TJ, Lawrence MG, Risbridger GP, Molloy P . . . Clark SJ. Critical evaluation of the Illumina MethylationEPIC BeadChip microarray for whole-genome DNA methylation profiling. *Genome Biol* 2016;**17**(1):208. https://doi.org/10.1186/s13059-016-1066-1.

41. Razin A. CpG methylation, chromatin structure and gene silencing-a three-way connection. *EMBO J* 1998;**17**(17):4905–8. https://doi.org/10.1093/emboj/17.17.4905.

42. Ritchie ME, Phipson B, Wu D, Hu Y, Law CW, Shi W, Smyth GK. limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic Acids Res* 2015;**43**(7). https://doi.org/10.1093/nar/gkv007 e47.

43. Robertson J, Robertson AB, Klungland A. The presence of 5-hydroxymethylcytosine at the gene promoter and not in the gene body negatively regulates gene expression. *Biochem Biophys Res Commun* 2011;**411**(1):40–3.

44. Skvortsova K, Zotenko E, Luu PL, Gould CM, Nair SS, Clark SJ, Stirzaker C. Comprehensive evaluation of genome-wide 5-hydroxymethylcytosine profiling approaches in human DNA. *Epigenet Chromatin* 2017;**10**:16. https://doi.org/10.1186/s13072-017-0123-7.

45. Society AC. Cancer facts & figures 2018. In. Atlanta: American Cancer Society; 2018.

46. Society AC. Cancer facts & figures 2019. In. Atlanta: American Cancer Society; 2019.

47. Tahiliani M, Koh KP, Shen Y, Pastor WA, Bandukwala H, Brudno Y . . . Rao A. Conversion of 5-methylcytosine to 5-hydroxymethylcytosine in mammalian DNA by MLL partner TET1. *Science* 2009;**324**(5929):930–5. https://doi.org/10.1126/science.1170116.

48. Vargas AJ, Harris CC. Biomarker development in the precision medicine era: lung cancer as a case study. *Nat Rev Cancer* 2016;**16**(8):525–37. https://doi.org/10.1038/nrc.2016.56.

49. Wang T, Guan W, Lin J, Boutaoui N, Canino G, Luo J . . . Chen W. A systematic study of normalization methods for Infinium 450K methylation data using whole-genome bisulfite sequencing data. *Epigenetics* 2015;**10**(7):662–9. https://doi.org/10.1080/15592294.2015.1057384.

50. Wu H, Wang C, Wu Z. A new shrinkage estimator for dispersion improves differential expression detection in RNA-seq data. *Biostatistics* 2013;**14**(2):232–43. https://doi.org/10.1093/biostatistics/kxs033.

51. Yang H, Liu Y, Bai F, Zhang J, Ma S, Liu J . . . Ye D. Tumor development is associated with decrease of TET gene expression and 5-methylcytosine hydroxylation. *Oncogene* 2013;**32**(5):663.

52. Yu M, Hon GC, Szulwach KE, Song CX, Zhang L, Kim A . . . He C. Base-resolution analysis of 5-hydroxymethylcytosine in the mammalian genome. *Cell* 2012;**149**(6):1368–80. https://doi.org/10.1016/j.cell.2012.04.027.

53. Zhou W, Laird PW, Shen H. Comprehensive characterization, annotation and innovative use of Infinium DNA methylation BeadChip probes. *Nucleic Acids Res* 2017;**45**(4):e22.