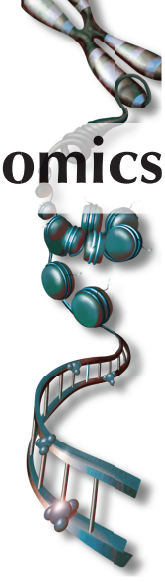
Editorial

For reprint orders, please contact: reprints@futuremedicine.com

Epigenomics



Gene body profiles of 5-hydroxymethylcytosine: potential origin, function and use as a cancer biomarker

Gerd P Pfeifer*,[1] & Piroska E Szabó[1]

[1]Center for Epigenetics, Van Andel Research Institute, Grand Rapids, MI, 49503, USA

*Author for correspondence: gerd.pfeifer@vai.org

"The DNA base 5-hydroxymethylcytosine (5hmC) strongly accumulates along transcribed sequences (gene bodies) of tissue-specific genes"

First draft submitted: 8 May 2018; Accepted for publication: 30 May 2018; Published online: 27 July 2018

Keywords: 5-hydroxymethylcytosine • cancer biomarkers • DNA methylation • gene bodies • TET enzymes • transcription

The DNA base 5-hydroxymethylcytosine (5hmC) strongly accumulates along transcribed sequences (gene bodies) of tissue-specific genes. We discuss the potential origin and function of 5hmC in gene bodies and describe how genomic 5hmC patterns may be useful in cancer diagnosis. Since 5-hydroxymethylcytosine (5hmC) was first described as a normally occurring DNA base in mammals in 2009 [1,2], many studies have investigated the mechanisms of its formation, established genomic profiles of 5hmC distribution in different tissues or cell lines and generated hypotheses to address its biological function. This DNA base is produced by enzymatic oxidation of 5-methylcytosine (5mC) at CpG dinucleotide sequences by a small family of three mammalian 5mC oxidases, the TET proteins, which use α-ketoglutarate, oxygen and Fe^{++} as cofactors [2,3]. The TET proteins have the ability to oxidize the initially formed 5hmC base further and sequentially produce 5-formylcytosine and 5-carboxylcytosine. The latter two bases can be removed from DNA by base excision repair initiated by thymine DNA glycosylase [4]. The completed pathway works as a mechanism to achieve active, replication-independent DNA demethylation leading to loss of methylation at CpG sites. However, given the considerable abundance of 5hmC in specific cell types (for example, it can reach a level of about 1% of all DNA cytosine bases in human brain-derived neurons), there has been the assumption that 5hmC has its own biological meaning and may be recognized by specific reader proteins [5]. Yet, these reader proteins have remained elusive with no clearly demonstrated roles *in vivo* and the function of 5hmC has remained enigmatic altogether.

5hmC genomic distribution

Methods to map the distribution of 5hmC along the genome include antibody-based immunoprecipitation methods (hMeDIP) followed by array or sequence analysis, pull-down approaches after specific derivatization of the hydroxymethyl groups by transfer of modified glucose residues and a few single base resolution approaches [6]. These profiling studies have shown that 5hmC is generally depleted in gene-poor areas of mammalian genomes but is enriched in transcribed sequences (gene bodies) and is also found close to intergenic and intragenic enhancer elements and at the edges of CpG-rich promoter regions [7]. ChIP-sequencing studies have shown that the 5hmC mark co-occurs with RNA polymerase II and the histone modifications H3K79me3 and H3K36me3, all markers of active transcription units, but is strongly depleted from regions associated with inactive chromatin such as those marked by the polycomb complex that produces H3K27me3. Of note, 5mC also rarely co-occurs with polycomb marks. As may be concluded from this data, 5hmC is clearly a gene-centered mark that correlates with active genes. At enhancer elements, 5hmC is most likely deposited by the TET2 5mC oxidase as shown by *Tet2*-specific or *Tet1/2/3* triple knockout experiments [8,9]. Both TET1 and TET3 contain a CXXC domain, a zinc finger module capable of binding to unmethylated CpG-rich DNA sequences. Therefore, it is thought that at least those TET1

Future Medicine

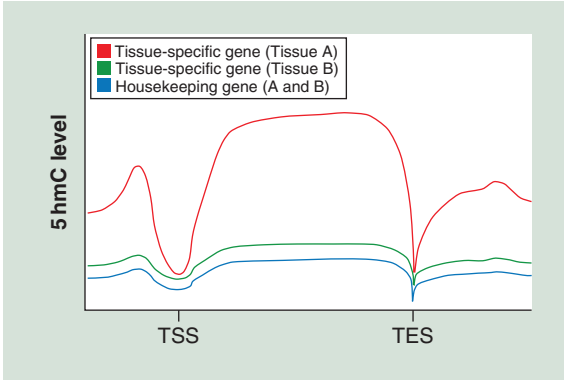


Figure 1. Schematic representation of a gene body profile of 5-hydroxymethylcytosine (5hmC).
The levels of 5hmC are very low near transcription start sites, increase throughout the gene bodies and then decline near the end of the transcription units (TES). This hypothetical gene is expressed in tissue A (red line) but not in tissue B (green line) with both tissues having similar global levels of 5hmC. A housekeeping gene expressed in all tissues is shown for comparison (blue line).

and TET3 isoforms that contain this domain are targeted to CpG islands, where they may function primarily by oxidizing 5mC bases that have been created there by erroneous *de novo* DNA methylation events [3,10]. The envisioned task of these CpG island-targeted TET proteins is to keep CpG islands free of DNA methylation.

Tissue-specificity of 5hmC gene body profiles

Within genes, 5hmC is depleted near the transcription start site and near the end of the transcription units but reaches much higher levels along the transcribed sequences (Figure 1). Many studies have shown that 5hmC in gene bodies positively correlates with gene expression levels [7]. Global and gene-specific levels of 5hmC are indicators of cellular state and tissue type [11]. Genes that are expressed at substantial levels in specific tissues or cell types will often carry higher levels of 5hmC when comparing with other tissues in which the same genes are not expressed but these differences also depend on the global levels of 5hmC in different cells. The observation that tissue-specifically expressed genes are marked by 5hmC has been made for example in several neuronal cell types, cardiomyocytes, various epithelial cells and immune cells. On the other hand, housekeeping genes, even though they may be expressed at high levels, contain 5hmC only at low frequencies. These genes include mitochondrial and ribosomal proteins, which despite being highly transcribed carry very low levels of 5hmC in their gene bodies [12]. It is unclear whether levels of 5hmC change within genes that undergo short-term induction, for example during the DNA damage response.

How are tissue-specific genes marked by 5hmC & what is the role of this DNA base within these genes?

TET proteins perform key roles in cell lineage differentiation processes [13]. Since 5hmC deposition seems to accompany upregulation of cell type-specific genes during the differentiation of cells from stem or progenitor cells, the question arises how this process occurs mechanistically. The simple explanation that TET proteins are constant companions of RNA polymerase II elongation complexes seems unlikely since housekeeping genes are poorly occupied by 5hmC. Since the genes most highly decorated by 5hmC, are tissue-specific genes, its occurrence in these genes must somehow be linked to the tissue specificity of their expression. Tissue specificity is usually determined by transcription factors that localize to promoter and/or enhancer regions of the genes that characterize a cell type. Therefore it is conceivable that TET proteins are recruited by tissue-specific transcription factors and, through unknown mechanisms, are then 'handed over' to the transcription elongation complexes to oxidize 5mC along gene bodies. In this context, it is of interest that TET2 has been identified as a binding partner of AF9, a component of transcriptional elongation complexes [14]. There are so far only relatively few examples where TET recruitment has been demonstrated in the form of direct physical interaction between TET proteins and a transcription factor. They include pluripotency factors, lineage-specific pioneer factors and other important regulators. Such interactions have been shown, for example for FOXA1 and TET1 [15], NANOG and TET1 [16], FOXO3A and TET2 [17], WT1 and TET2 and TET3 [18] and the master hematopoietic transcription factor RUNX1 and TET2 and TET3 [19]. Pioneer factors operate to open inaccessible chromatin during lineage commitment and cell differentiation processes. However, pioneer factors are highly specialized and there are many other cell type-specific transcription factors that determine the full repertoire of tissue-specific gene expression often following the initial activity of pioneer factors. Given the fact that mammalian organisms have in the order of 200 different cell types and hence at least that many transcription factor combinations that determine cellular identity, it

is difficult to imagine how only three TET proteins would be able to interact with so many transcription factors in a specific manner. The question is if differences may exist in chromatin structure, promoter composition and/or cis-regulatory elements of tissue-specific genes compared with housekeeping genes that may specify TET occupancy, perhaps in a combinatorial manner. Curiously, *Drosophila* and other organisms that have very low levels of or completely lack 5mC and 5hmC also have tissue-specific genes that function in the absence of 5hmC-dependent (and 5mC-dependent) gene control mechanisms. This raises the question as to whether 5hmC is needed for high fidelity gene control only in long-living creatures.

One other unresolved issue is if and how 5hmC in gene bodies increases transcript levels. Genetic ablation of the DNA methyltransferase *Dnmt3b*, which causes depletion of 5mC and consequentially 5hmC in gene bodies, leads to increased transcriptional noise due to intragenic aberrant transcription initiation [20]. Simultaneous inactivation of the *Tet1* and *Tet3* genes also resulted in loss of transcriptional fidelity in early embryos [21]. It is possible that 5hmC is even more potent than 5mC in suppressing inappropriate transcription initiation within gene bodies. However, formal proof for this concept as well as an understanding of the mechanisms of this regulation are so far lacking.

Low levels of 5hmC in cancer

The 5hmC base is strongly depleted in almost all types of human cancer in comparison with the normal tissues in which the tumors emerge [22]. It is still unclear how this depletion arises and what the consequences are with regards to tumorigenesis. However, it is well known that malignant tumors lose many features of normal tissue/cell architecture and become de-differentiated. When viewing 5hmC as a marker of cell- or tissue-specific genes, its loss in cancer cells would appear as a logical event. However, it remains unknown if loss of 5hmC in gene bodies is a cause or a consequence of malignant transformation and whether other factors contribute to its loss [23].

5hmC profiles as a cancer marker in liquid biopsies

The global loss of 5hmC in tumors may be accompanied by gain of this mark in certain cancer-specific genes. Therefore, characterization of 5hmC profiles has emerged as an attractive practical approach for tumor classification and tumor detection. Interestingly, these altered patterns of 5hmC are not only detectable in tumor-normal tissue pairs obtained from tumor resections or biopsies, but can be identified in circulating free DNA from cancer patients. For example, 5hmC-based liquid biomarkers distinguishing healthy controls from individuals with cancer were highly predictive for colorectal and gastric cancers [24] and for lung cancer, hepatocellular carcinoma and pancreatic cancer [25]. Although tumor-specifically methylated DNA has previously been used as a tumor marker for blood-based biopsies, most sodium bisulfite-based DNA methylation detection methods further degrade the already highly fragmented DNA in blood/serum samples. However, the selective labeling approach used for 5hmC profiling does not include bisulfite treatment and may therefore have advantages over previously used epigenetic assays to detect cancer in serum samples. These promising approaches will need to be further optimized and applied to larger clinical studies.

Financial & competing interests disclosure

Work of the authors was supported by NIH grants CA160965 (to GP Pfeifer) and GM064378 (to PE Szabó). The authors have no other relevant affiliations or financial involvement with any organization or entity with a financial interest in or financial conflict with the subject matter or materials discussed in the manuscript apart from those disclosed.

No writing assistance was utilized in the production of this manuscript.

References

1. Kriaucionis S, Heintz N. The nuclear DNA base 5-hydroxymethylcytosine is present in Purkinje neurons and the brain. *Science* 324(5929), 929–930 (2009).
2. Tahiliani M, Koh KP, Shen Y *et al.* Conversion of 5-methylcytosine to 5-hydroxymethylcytosine in mammalian DNA by MLL partner *TET1*. *Science* 324(5929), 930–935 (2009).
3. Rasmussen KD, Helin K. Role of TET enzymes in DNA methylation, development and cancer. *Genes Dev.* 30(7), 733–750 (2016).

4. He YF, Li BZ, Li Z *et al.* Tet-mediated formation of 5-carboxylcytosine and its excision by thymine DNA glycosylase in mammalian DNA. *Science* 333(6047), 1303–1307 (2011).
5. Song J, Pfeifer GP. Are there specific readers of oxidized 5-methylcytosine bases? *Bioessays* 38(10), 1038–1047 (2016).
6. Wu X, Zhang Y. TET-mediated active DNA demethylation: mechanism, function and beyond. *Nat. Rev. Genet.* 18(9), 517–534 (2017).
7. Wen L, Tang F. Genomic distribution and possible functions of DNA hydroxymethylation in the brain. *Genomics* 104(5), 341–346 (2014).
8. Hon GC, Song CX, Du T *et al.* 5mC oxidation by *Tet2* modulates enhancer activity and timing of transcriptome reprogramming during differentiation. *Mol. Cell* 56(2), 286–297 (2014).
9. Lu F, Liu Y, Jiang L *et al.* Role of Tet proteins in enhancer activity and telomere elongation. *Genes Dev.* 28(19), 2103–2119 (2014).
10. Jin SG, Zhang ZM, Dunwell TL *et al. Tet3* reads 5-carboxylcytosine through Its CXXC domain and is a potential guardian against neurodegeneration. *Cell Rep.* 14(3), 493–505 (2016).
11. Laird A, Thomson JP, Harrison DJ *et al.* 5-hydroxymethylcytosine profiling as an indicator of cellular state. *Epigenomics* 5(6), 655–669 (2013).
12. Lin IH, Chen YF, Hsu MT. Correlated 5-hydroxymethylcytosine (5hmC) and gene expression profiles underpin gene and organ-specific epigenetic regulation in adult mouse brain and liver. *PLoS ONE* 12(1), e0170779 (2017).
13. Scott-Browne JP, Lio CW, Rao A. TET proteins in natural and induced differentiation. *Curr. Opin. Genet. Dev.* 46 202–208 (2017).
14. Qiao Y, Wang X, Wang R *et al.* AF9 promotes hESC neural differentiation through recruiting *TET2* to neurodevelopmental gene loci for methylcytosine hydroxylation. *Cell Discov.* 1, 15017 (2015).
15. Yang YA, Zhao JC, Fong KW *et al. FOXA1* potentiates lineage-specific enhancer activation through modulating *TET1* expression and function. *Nucleic Acids Res.* 44(17), 8153–8164 (2016).
16. Costa Y, Ding J, Theunissen TW *et al.* NANOG-dependent function of *TET1* and *TET2* in establishment of pluripotency. *Nature* 495(7441), 370–374 (2013).
17. Li X, Yao B, Chen L *et al.* Ten-eleven translocation 2 interacts with forkhead box O3 and regulates adult neurogenesis. *Nat. Commun.* 8, 15903 (2017).
18. Rampal R, Alkalin A, Madzo J *et al.* DNA hydroxymethylation profiling reveals that *WT1* mutations result in loss of *TET2* function in acute myeloid leukemia. *Cell Rep.* 9(5), 1841–1855 (2014).
19. Suzuki T, Shimizu Y, Furuhata E *et al.* RUNX1 regulates site specificity of DNA demethylation by recruitment of DNA demethylation machineries in hematopoietic cells. *Blood Adv.* 1(20), 1699–1711 (2017).
20. Neri F, Rapelli S, Krepelova A *et al.* Intragenic DNA methylation prevents spurious transcription initiation. *Nature* 543(7643), 72–77 (2017).
21. Kang J, Lienhard M, Pastor WA *et al.* Simultaneous deletion of the methylcytosine oxidases *Tet1* and *Tet3* increases transcriptome variability in early embryogenesis. *Proc. Natl Acad. Sci. USA* 112(31), E4236–E4245 (2015).
22. Jin SG, Jiang Y, Qiu R *et al.* 5-Hydroxymethylcytosine is strongly depleted in human cancers but its levels do not correlate with *IDH1* mutations. *Cancer Res.* 71(24), 7360–7365 (2011).
23. Pfeifer GP, Jin SG. Defective 5-methylcytosine oxidation in tumorigenesis. *Encyclopedia of Cancer* (*3rd Edition*) Elsevier (2018) (In press).
24. Li W, Zhang X, Lu X *et al.* 5-Hydroxymethylcytosine signatures in circulating cell-free DNA as diagnostic biomarkers for human cancers. *Cell Res.* 27(10), 1243–1257 (2017).
25. Song CX, Yin S, Ma L *et al.* 5-Hydroxymethylcytosine signatures in cell-free DNA provide information about tumor types and stages. *Cell Res.* 27(10), 1231–1242 (2017).