Author Manuscript

Author Manuscript

Author Manuscript

Author Manuscript

HHS Public Access

Author manuscript

Nat Struct Mol Biol. Author manuscript; available in PMC 2015 December 07.

Published in final edited form as:

Nat Struct Mol Biol. 2015 September ; 22(9): 656–661. doi:10.1038/nsmb.3071.

Charting oxidized methylcytosines at base resolution

Hao Wu[1,2,3,4] and **Yi Zhang**[1,2,3,4,#]

[1]Howard Hughes Medical Institute

[2]Program in Cellular and Molecular Medicine, Boston Children's Hospital, Boston, MA 02115

[3]Department of Genetics, Harvard Medical School, WAB-149G, 200 Longwood Avenue, Boston, MA 02115

[4]Harvard Stem Cell Institute, Harvard Medical School, WAB-149G, 200 Longwood Avenue, Boston, MA 02115

Abstract

DNA cytosine methylation (5-methylcytosines) represents a key epigenetic mark and is required for normal development. Iterative oxidation of 5mC by TET family of DNA dioxygenases generates three oxidized nucleotides, 5-hydroxymethylcytosine (5hmC), 5-formylcytosine (5fC), and 5-carboxycytosine (5caC), in the mammalian genome. Recent advances in genomic mapping techniques for these oxidized bases suggest that 5hmC/5fC/5caC are not only functionally relevant to the process of active reversal of 5mC, but may also possess unique regulatory functions. This perspective highlights the potential gene regulatory functions of these oxidized cytosine bases in the mammalian genome, and discusses the principles and limitations of recently developed base-resolution mapping technologies.

Introduction

Methylation of DNA allows the genome to carry regulatory information beyond its canonical function as genetic blueprint. In bacteria, methylation can occur on either adenosine or cytosine, impacting diverse biological processes such as degrading foreign DNA, tracking mismatch repair and regulating DNA replication [1]. DNA methylation at the 5-position of cytosine (5mC) is evolutionarily conserved in many eukaryotic organisms and has been functionally linked to gene expression regulation and genome integrity maintenance [2]. Intriguingly, recent studies indicate that adenosine methylation such as N6-methyladenision (6mA), is present in eukaryotic organisms (*C. elegans* and *D. melanogaster*) that are previously thought to lack DNA methylation [3,4], raising the possibility that DNA methylation has a general role in eukaryotic biology.

In vertebrates, cytosine methylation (5mC) is the predominant form of DNA modification and occurs throughout the entire genome, suggesting that their methylation might be a default state [5,6]. *De novo* DNA methyltransferases (DNMT3A and DNMT3B) primarily target 5mC to palindromic CpG dinucleotide and the maintenance DNA methyltransferase

[#]To whom correspondence should be addressed: Phone: (617) 713-8666, Fax: (617) 713-8665, yzhang@genetics.med.harvard.edu.

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

(DNMT1) enables faithful propagation of CpG methylation patterns through cell divisions [7]. Heritable CpG methylation (mCpG) is therefore considered as a classic 'epigenetic' mark and is believed to be functionally involved in many forms of long-term epigenetic memory processes, such as genomic imprinting, X chromosome inactivation and silencing of repeats [8]. Interestingly, highly dynamic changes of DNA methylation take place at a genome-wide scale during early embryonic development and is required for critical biological processes such as erasure of parental-origin-specific imprints in developing primordial germ cells (PGCs) [9,10]. In addition, genome-wide mapping of 5mC revealed that active gene regulatory sequences, such as gene promoters and distal enhancers, are hypomethylated [11,12]. Because these DNA demethylation processes are not always coupled with DNA replication-dependent passive dilution of 5mC, specific enzymatic activity may exist for active removal of 5mC in vertebrates. Recent discovery of the ten-eleven translocation (TET) family of 5mC dioxygenases has provided a biochemically plausible pathway for catalyzing active DNA demethylation process [13,14]. TET proteins convert 5mC into 5-hydroxymethylcytosine (5hmC) [15–17]. Further successive oxidations mediated by TET result in 5-formylcytosine (5fC) and 5-carboxylcytosine (5caC) [18,19], both of which can be efficiently excised by Thymine DNA glycosylase (TDG) and restored to unmodified cytosines through base excision repair (BER) pathway [18,20,21]. Genetic studies of TET mutant mice indicated that these 5mC oxidases play important roles in diverse biological processes, including embryonic development, stem cell differentiation, erasure of genomic imprinting, learning and memory, and cancer [reviewed in [13,14]].

In addition to function as intermediates of an active DNA demethylation pathway (Fig. 1a), growing evidence indicates that these oxidized methylcytosines may possess unique regulatory functions. To gain insight into the potential function of 5hmC, 5fC, and 5caC, several studies have attempted to identify reader proteins for oxidized methylcytosines [22,23] (Fig. 1b). These studies not only identified proteins that are functionally linked to DNA repair process, but also uncovered transcription factors and chromatin modifying enzymes as candidate reader proteins for oxidized 5mC bases. Interestingly, the number of identified candidates for 5fC and 5caC is much higher than that of 5hmC, possibly due to the unique chemical properties of formyl and carboxyl groups of these two highly oxidized bases. In addition, biochemical and structural evidences indicate that 5fC and 5caC within the gene body may reduce the elongation rate of RNA polymerase II (Pol II) [24,25] (Fig. 1c). Furthermore, biophysical studies suggest that these oxidized bases may have impact on base-pairing and DNA structure [26,27], suggesting that these chemical modifications may affect DNA-templated processes by directly impact DNA conformation. Lastly, potential single or double strand breaks associated with the DNA repair process downstream of 5fC/5caC excision may contribute to gene regulation [13].

Understanding the mechanisms underlying these roles require the ability to comprehensively profile the distribution of the reactions that TET and TDG enzymes catalyze in the mammalian genome. Recent technological advances have results in genomic maps of oxidized 5mC bases (5hmC/5fC/5caC) at unprecedented resolution, revealing that TET-mediated 5mC oxidation events are preferentially targeted to genomic regions associated with gene regulatory functions. Despite these intriguing but correlative observations, exactly how oxidized 5mC bases exert their function at these regulatory regions is largely unclear

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

and is under active investigation [28]. This perspective summarizes recent advances in genomic mapping methods for oxidized 5mC bases (5hmC/5fC/5caC), and discuss the potential functions of 5mC and its oxidized derivatives as stable and transient epigenetic marks in gene regulation.

Genomic mapping of oxidized 5-methylcytosines at single-base resolution

As the first enzymatic product of TET-mediated 5mC oxidation, 5hmC is detected in a broad spectrum of mammalian tissues. In contrast to the relatively high 5mC levels, which are stable across somatic tissues (~4% of total cytosines), the levels of 5hmC are highly variable in a tissue specific manner. For instance, 5hmC can be as high as 40% of the 5mC level in Purkinje neurons in cerebellum [15], ~5% of the 5mC level in embryonic stem cells (ESCs) [16,17], and as low as ~1% of 5mC in some immune cells [29]. While 5hmC is generally considered as an intermediate of TET-mediated active DNA demethylation, recent studies have suggested that the majority of 5hmC is a stable modification in mouse tissues [30]. Iterative oxidation of 5hmC by TET generates 5fC and 5caC, which are present ~100-fold lower than the level of 5hmC in wild-type ESCs [19]. The low abundance of 5fC [~20 Parts Per Million (ppm) cytosines in ESCs] and 5caC (~3 ppm) is in part due to the robust 5fC/5caC excision activity of TDG in mammalian cells. Indeed, depletion of TDG in mouse ESCs results in 5–10 fold increase in the levels of 5fC and 5caC [31,32], suggesting that the majority of 5fC and 5caC are only transiently present in the mammalian genome. One potential exception to the transitory presence of 5fC and 5caC is the initial generation and eventual replication-dependent dilution of 5fC and 5caC in early pre-implantation embryos [33], where TDG mRNAs are not detected. Using a stable isotope tracing strategy, a recent study indicates that 5fC can be a stable DNA modification in nonproliferating cells or postnatal tissues [34], albeit at low levels (ranging from 0.2 to 15 ppm in different tissues). The total amount of oxidized cytosine bases can be quantified by several methods, including thin layer chromatography, modification-specific antibodies, chemical tagging, and mass spectrometry [reviewed in [35]].

Genome-wide mapping of oxidized 5mC variants in mammalian genomes is technically challenging due to their extremely low abundances. Early efforts in mapping these modified nucleotides relied predominantly on affinity-enrichment based methods that utilize either modification-specific antibodies or chemical tagging [36,37]. However, genomic maps of 5hmC/5fC/5caC generated by these affinity-enrichment methods are of limited resolution (several hundred base-pairs), represent only relative enrichment (over control experiments) and lack precise strand distribution information. To address these limitations, methods have been developed recently to discriminate the different 5mC oxidation variants at single-base resolution at genome-scale (Fig. 2). Importantly, results from these studies have provided new insights into the biological functions of these oxidized cytosine bases in the mammalian genome.

Base-resolution mapping of 5hmC in mammalian genomes

The current method of choice for mapping cytosine methylation at single-base resolution is sodium bisulfite ($NaHSO_3$) conversion of genomic DNA followed by next-generation sequencing (BS-seq). In BS-seq, 5mC and 5hmC are resistant to deamination and

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

consequently be read out as cytosine (C) after PCR amplification; unmodified C, 5fC and 5caC are deaminated in bisulfite treatment and read out as thymine (T) in subsequent sequencing. Thus, methylation signals (C) in standard BS-seq represent the sum of 5mC and 5hmC. To map 5hmC at base-resolution, several modified bisulfite-sequencing method have been developed (Fig. 2).

The first method, termed oxidative bisulfite sequencing (oxBS-seq), utilizes potassium perruthenate ($KRuO_4$) to specifically oxidize 5hmC to 5fC. Thus, 5hmC (now converted to 5fC) is detected as T, whereas only 5mC remains intact and is read out as C. While 5mC is directly mapped, subtracting signals of oxBS-seq maps (5mC) from those of traditional BS-seq (5mC+5hmC) would reveal the absolute level and precise location of 5hmC [38]. Because this method requires subtraction between two random sampling-based BS-seq experiments, it requires very deep sequencing coverage to achieve relatively high-confidence mapping of 5hmC. One strategy to reducing sequencing effort is to combine reduced representation BS-seq (RRBS) strategy with oxBS-seq [38]. In this oxRRBS approach, converted DNA was first digested with the *MspI* enzyme (C^CGG, "^" is the cutting site) to enrich for CpG-containing DNA fragments. It thus allows deep (~120× per cytosine) and selective sequencing of a fraction of the genome that is highly enriched for CpG-rich sequences such as CpG islands (CGIs) and certain-types of repeats. Application of the oxRRBS to mouse ESCs have revealed that 5hmC is relatively enriched at transcriptionally poised and CpG-rich promoters that drive expression of lineage-specific transcription factors during cellular differentiation. In fact, 5hmC enrichment at promoter is negatively correlated with the steady state level of transcription. These results are consistent with results from 5hmC maps generated by affinity enrichment-based methods [39,40].

The second base-resolution 5hmC mapping method is called Tet-assisted bisulfite sequencing (TAB-seq), which involves TET-mediated enzymatic conversion of non-5hmC modifications (5mC/5fC) to 5caC followed by bisulfite sequencing [41]. In TAB-seq, 5hmC is first protected from TET-mediated oxidation by glucosylation using β-glucosyltransferase (β-GT). 5mC and 5fC are then oxidized to 5caC in the presence of high concentration of recombinant TET1 proteins. Thus, C/5mC/5fC/5caC are detected as T, whereas only glucosylated 5hmC is read as C. A major advantage of TAB-seq is that 5hmC is directly mapped, which reduces the sequencing effort needed for high confidence 5hmC mapping [41]. As a result, application of TAB-seq to human and mouse ESCs enabled the generation of first whole-genome maps of 5hmC at base-resolution. With a medium sequencing depth (17–26× per cytosine) and stringent statistical filtering, TAB-seq can identify the precise position and quantify the absolute abundance of 5hmC of 20% or higher. A total of ~0.7 and ~2.1 million high-confidence 5hmC sites have been identified in human and mouse ESCs, respectively (Table 1). Whole-genome map of 5hmC revealed that 5hmC is highly enriched at many CpG-poor and promoter-distal gene regulatory regions such as p300-marked enhancers and CTCF-bound insulators, which are generally under-represented in RRBS datasets. Previous affinity enrichment-based maps also suggested that 5hmC is enriched at distal *cis*-regulatory elements [39,42,43], but lacked the resolution required for investigating the relationship between 5hmC position and transcription factor (TF) binding motifs. The base-resolution map revealed that 5hmC is typically not detected within TF binding sites but

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

rather is most enriched in regions immediately adjacent to TF motifs, generating bimodal peaks of 5hmC centered at the motif. Mammalian nervous systems possess the highest level of 5hmC among all somatic tissues. Age-dependent increase of total 5hmC levels in cortex, hippocampus and cerebellum supports a role of 5hmC as a stable epigenetic mark in neuronal genomes [44]. Indeed, whole-genome TAB-seq analysis of fetal and young adult frontal cortex not only confirmed a general increase of 5hmC at various genomic features in adult brain compared to fetal brains, but also revealed an overall positive correlation between intragenic 5hmC enrichment (in CpG context) and transcriptional activity [45]. In addition to ESCs and brains, TAB-seq has also been used to analyze the 5hmC pattern in two-cell embryos [46]. Combined with paternal/maternal allele specific analysis, TAB-seq (18× per cytosine) has identified ~0.1 million 5hmCpGs in paternal genome and ~0.12 million 5hmCpGs in maternal genome, suggesting that TET-mediated 5mC oxidation takes place on both sperm- and oocyte-derived chromosomes. Indeed, genetics analysis of wild-type and Tet3 maternal knockout (KO) embryos indicates that Tet3 deficiency affect DNA demethylation of both paternal and maternal genomes in one-cell zygotes [47,48].

The low level of 5hmC in the genome suggests that base-resolution mapping methods such as oxBS-seq and TAB-seq, require ultra-deep sequencing coverage (>100x) to achieve high-confidence identification of 5hmC sites of low abundance. To overcome this limitation, a recent study has developed a more sensitive approach using the PvuRst1l family of 5hmC-dependent restriction endonucleases [49]. PvuRts1l endonuclease recognizes 5hmC and creates a double-stranded break 11–12 bp downstream, leaving a 2-bp 3′ overhang. The resulting DNA fragments can then be converted into sequencing libraries and the location of 5hmC is determined by mapping the cleavage sites. Combining chemical labeling-based 5hmC enrichment method, hMe-Seal [50], with PvuRst1l digestion, the method (termed Pvu-Seal-seq) allows cost-efficient gnome-wide 5hmC mapping at base-resolution. Using Pvu-Seal-seq, the authors have identified 20.8 million reproducible 5hmC sites in mouse ESCs (present in two technical replicates) (Table 1), 10-time higher than the number of 5hmC sites identified by whole-genome TAB-seq analysis. Among 5hmC sites identified by Pvu-Seal-seq, 24% of them were in non-CpG context (~50% in CpA). This observation is surprising as results from both oxRRBS and whole-genome TAB-seq suggest that nearly all 5hmC sites are detected in the CpG context (~99%) [38,41]. Interestingly, while 64% of 5hmCpG sites are conserved between the two biological replicates, only 24% of 5hmC in non-CpG context (5hmCpH) are reproducibly detected by Pvu-Seal-seq. The results imply that 5hmCpH is much less stable than 5hmCpG, possibly due to the fact that 5mC in non-CpG context cannot be faithfully maintained during cell division. Locus-specific TAB-seq analysis (sequencing depth >100x) confirmed some of the newly identified 5hmCpH sites, and revealed that the average 5hmC level at non-CpG context (2.8%) is significantly lower than that of 5hmCpG (11.4%), highlighting the need of ultra-high sequencing depth for detecting 5hmCpH with high confidence. Overall, these results suggest that Pvu-Seal-seq has enhanced sensitivity and can detect low abundance and/or unstable 5hmC sites in the genome without the need of ultra-high sequencing depth. However, due to the initial step of affinity-enrichment, Pvu-Seal-seq cannot quantify the absolute percentage of 5hmC level. Thus, it can only be used to quantify relative changes of 5hmC levels.

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

Base-resolution mapping of 5fC and 5caC in mammalian genomes

Available evidence suggests that oxidative modification of 5mC by TET proteins promotes DNA demethylation by either replication-dependent dilution of 5hmC (by impeding DNMT1 function) or TDG-mediated 5fC/5caC excision followed by BER [13]. The active DNA demethylation pathway involving generation and excision repair of 5fC/5caC is of particular interest as it may occur in both proliferating and post-mitotic cells. The observation that 5hmC in CpG context is relatively stable *in vivo* suggests that identifying methylated CpGs that are targeted for active DNA demethylation requires methods that permit quantitative measurement of 5fC and 5caC levels at single-base resolution. Despite the scarcity of 5fC and 5caC in the genome, several modified BS-seq methods have been developed for mapping 5fC and 5caC at single-base resolution (Fig. 2).

Subtraction-dependent methods—This group of base-resolution mapping methods utilizes specific chemical treatment to protect 5fC or 5caC from deamination by bisulfite treatment. To map 5fC, fCAB-seq (5fC chemically assisted bisulfite sequencing) utilizes O-ethylhydroxylamine ($EtONH_2$) to protect 5fC [32]. In a similar approach called reduced bisulfite sequencing (redBS-seq), 5fC is selectively reduced by sodium borohydride ($NaBH_4$) to 5hmC [51]. To protect 5caC, caCAB-seq (5caC chemically assisted bisulfite sequencing) takes advantage of 1-ethyl-3-[3-dimethylaminopropyl] carbodiimide hydrochloride (EDC)-based coupling to chemically block 5caC from deamination [52]. However, all three methods require subtracting signals of standard BS-seq from those of modified BS-seq to determine the position and abundance of 5fC (fCAB-seq and redBS-seq) or 5caC (caCAB-seq). The low abundance of 5fC/5caC, coupled with the possibility that chemicals used in these methods may react with moieties (e.g. $EtONH_2$ reacts with abasic sites or unmodified C) on DNA other than their intended targets (5fC or 5caC) may complicate the interpretation of the results. To reduce sequencing efforts, various enrichment strategies were integrated with subtraction-dependent 5fC/5caC base-resolution mapping methods to generate genome-scale maps. By combining chromatin immunoprecipitation (using antibody against mono-methylation of lysine 4 on histone H3 [H3K4me1]) with fCAB-seq (termed H3K4me1-fCAB-seq), 5fC abundance can be examined within a fraction of genome that are associated with active/poised enhancer activity [32]. Integrating redBS-seq and RRBS allows 5fC mapping within genomic regions containing CpG-rich sequences (mostly gene promoters) [51]. It is important to point out that the *MspI* enzyme used in standard RRBS only partially digests 5fC containing C^CGG sites and completely fail to cut 5caCpGs [19]. Thus, either reducing 5fC to 5hmC before *MspI* digestion or choosing an enzyme (e.g. TaqαI, T^CGA) less influenced by 5fC/5caC modifications should be considered when using the RRBS strategy for 5fC/5caC base-resolution analysis. More recently, DNA immunoprecipitation (using antibody against 5fC or 5caC) is used to first enrich DNA fragments containing these rare modified bases before fCAB-seq and caCAB-seq assays. In this DIP-CAB-seq strategy, although the position of 5fC and 5caC can be precisely mapped, only relative enrichment of 5fC or 5caC can be determined [53]. Because of the need of ultra-high sequencing depth, subtraction-based 5fC/5caC mapping methods are not well-suited for genome-scale quantitative analysis of 5fC and 5caC at base-resolution.

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

Subtraction-independent methods—Because 5fC and 5caC are present at extremely low levels in the genome, it is highly desirable to map these rare bases without the need of subtracting two BS-seq experiments. In addition, as both 5fC and 5caC are substrate for TDG-mediated excision repair, determine the strand-specific preference of TET/TDG-mediated demethylation activity requires the ability to simultaneously map both 5fC and 5caC in a single experiment. To circumvent these limitations, three groups recently developed a modified BS-seq method, called methylase-assisted bisulfite sequencing (MAB-seq), to directly map 5fC/5caC at single-base resolution [48,54,55]. In MAB-seq, genomic DNA is first treated with the bacterial DNA CpG methyltransferase *M.Sss*I, an enzyme that is known to efficiently methylate CpG dinucleotides. Bisulfite conversion of methylase-treated DNA may lead to deamination of only 5fC and 5caC; originally unmodified CpGs are protected as 5mCpG. Subsequent sequencing would reveal 5fC and 5caC as T, whereas C/5mC/5hmC would be read out as C. Notably, MAB-seq is unable to distinguish 5fC/5caC from unmodified C within a non-CpG context due to the poor activity of *M.Sss*I towards CpH. Building upon MAB-seq, we have further developed a method termed caMAB-seq (5caC methylase-assisted bisulfite sequencing) to directly map 5caC at single-base resolution. This modified version of MAB-seq takes advantage of the fact that 5fC can be selectively reduced by $NaBH_4$ to 5hmC. After *M.Sss*I-treated DNA was further incubated with $NaBH_4$, only 5caC was read out as T; original 5fC, along with C/5mC/5hmC, is read as C in caMAB-seq. By integrating $NaBH_4$-mediated 5fC reduction into the Pvu-Seal-seq workflow, a modified Pvu-Seal-seq method can also directly detect 5fC at base-resolution.

Genomic distribution of 5fC and 5caC in stem cells and early development—The whole-genome base-resolution map of 5fC was first generated in two-cell embryos using fCAB-seq assays [46]. By subtracting standard BS-seq signals from those of fCAB-seq, this analysis identified ~0.95 million 5fCpGs genome-wide with an averaged level of ~50% at individual CpGs. Given the moderate sequencing depth (18× per cytosines) and relatively conservative statistical filtering strategy (Fisher's exact test), the number of 5fCpGs present in the genome of two-cell embryos is likely under-estimated. Sanger-sequencing-based locus-specific MAB-seq was also used to analyze early developing embryos, and this analysis did not detect substantial levels of 5fC/5caC at several CpG-rich gene promoters [48].

Whole-genome base-resolution map of 5fC/5caC has recently been generated using MAB-seq for wild-type and *Tdg*-depleted mouse ESCs [54,55]. Deep sequencing analysis (>50× per cytosine) of wild-type mouse ESCs has identified ~0.3 million 5fC/5caC-modified CpGs (5fC/5caCpGs) with an average modification level of 8–10% [55] (Table 1). Interestingly, 5fC/5caCpGs are found to be enriched at active gene promoters and enhancers in wild-type ESCs, co-localizing with both TET1/2 and TDG proteins [55]. Reduced representation version of MAB-seq (RRMAB-seq) allows more focused analysis of 5fC/5caC at CpG-rich promoters in wild-type and Tdg-depleted cells, revealing a Dnmt/TET/TDG-coupled DNA methylation/demethylaion dynamic process at actively transcribed gene promoters. Because both the increase of 5mC in *Tet1/2*-depleted cells (from ~2% to ~5%) and the increase of 5fC/5caC level in *Tdg*-depleted cells are quite modest at these active promoters in mouse ESCs, TET/TDG-mediated active DNA demethylation pathway may only play a relatively minor role in protecting these CpG-rich promoters from being hypermethylated. Redundant

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

mechanisms such as H3K4 methylation-dependent repulsion of DNMTs [56] and KDM2B (also known as FBXL10 or CXXC2) mediated protection against DNA hypermethylation [57] may compensate for the loss of function of TET enzymes.

Using *Dnmt* triple KO and *Tet1/2* double KO mouse ESCs (5fC/5caC-free genome) as the reference for controlling false-discovery rate, whole-genome base-resolution 5fC/5caC mapping (14× per cytosine) in *Tdg*-depleted mouse ESCs identified ~0.7 million 5fC/5caCpGs with an average of 20–30% modification level [54] (Table 1). 5fC/5caCpGs are most enriched at DNase I hypersensitivity sites, H3K4me1-marked ESC enhancers, CTCF-bound insulators and exonic regions. Strand-specific comparative analysis of 5fC/5caCpG and 5hmCpG base-resolution mapping revealed that the majority of 5hmCpGs (>90%) do not overlap with 5fC/5caCpGs, suggesting proteins exhibit distinct catalytic processivity at different CpGs. Further analysis suggests that the processivity of TET-mediated iterative oxidation positively correlates with local chromatin accessibility, such that 5fC/5caCpGs (including both 5fC/5caC-alone and dual modified CpGs) are associated with higher levels of DNase I hypersensitive sites, histone variants (H2A.Z and H3.3) known to destabilize nucleosome structure, and pluripotency-related TFs (Oct4/Nanog/Sox2) than 5hmC-alone CpGs [54]. Similar to strand asymmetry of 5hmCpGs [41], 5fC/5caCpGs also exhibit strong preference for asymmetrical modification [54,55], suggesting that single-strand breaks (SSBs) rather than double-strand breaks (DSBs) are the predominant intermediates during TET/TDG-mediated active DNA demethylation process. While 5hmC in the non-CpG context is rare and/or unstable, the presence of low abundance 5fC/5caCpGs cannot be excluded. Indeed, base-resolution 5fC mapping by the sensitive Pvu-Seal-seq has identified 6.2 million 5fC sites (75% in CpG and 25% in CpH) in wild-type mouse ESCs [49] (Table 1), significantly higher than the number of 5fC/5caC sites identified by MAB-seq. Future studies are needed to resolve this discrepancy between different methods.

Concluding remarks

The discovery of oxidized forms of methylcytosines, including 5hmC, 5fC and 5caC, in mammalian genomes has stimulated intense effort to map and quantify these modifications in different cell types, particularly in embryonic stem cells and developing brain tissues. The major challenge for such efforts is the scarcity of the 5mC oxidation products in mammalian cells. Application of BS-seq on chemically or enzymatically modified genomic DNA has generated single-base resolution maps of 5hmC, 5fC and 5caC. Comparative analysis of these maps with other epigenomic maps suggests that 5hmC/5fC/5caC are not randomly distributed in the mammalian genome, but rather preferentially enriched at specific genomic features with important gene regulatory functions.

In conclusion, recent technological advances in single-base resolution profiling of oxidized methylcytosine bases have provided significant new insights into the mechanism and function of TET- and TDG-mediated active DNA demethylation process. However, new experimental methods and strategies are needed to allow us to achieve integrated analysis of all five distinct cytosine modification states in small number of cells, particularly in preimplantation embryos and primordial germ cells where dynamic global DNA methylation changes are observed. Emerging technologies such as SMRT and nanopore sequencing have

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

the potential for direct reading of DNA modifications on single molecules without the need of DNA amplification or bisulfite conversion [36,37] (Table 1). However, the throughput and accuracy of these third-generation sequencing methods need to be substantially improved before they can be used for analyzing complex mammalian genomes.

Acknowledgments

We thank L.M. Tuesta for critical reading of the manuscript. We apologize to colleagues whose work cannot be cited owing to space constrains. This work was supported by NIH grants GM68804 and U01DK089565 (to Y.Z.). H.W. was supported by a postdoctoral fellowship from the Jane Coffin Childs Memorial Fund for Medical Research and is currently supported by the National Human Genome Research Institute (K99HG007982). Y.Z. is an Investigator of the Howard Hughes Medical Institute.

References

1. Arber W, Dussoix D. Host specificity of DNA produced by Escherichia coli. I. Host controlled modification of bacteriophage lambda. J Mol Biol. 1962; 5:18–36. [PubMed: 13862047]
2. Law JA, Jacobsen SE. Establishing, maintaining and modifying DNA methylation patterns in plants and animals. Nat Rev Genet. 2010; 11:204–20. [PubMed: 20142834]
3. Greer EL, et al. DNA Methylation on N(6)-Adenine in C. elegans. Cell. 2015; 161:868–78. [PubMed: 25936839]
4. Zhang G, et al. N(6)-methyladenine DNA modification in Drosophila. Cell. 2015; 161:893–906. [PubMed: 25936838]
5. Suzuki MM, Bird A. DNA methylation landscapes: provocative insights from epigenomics. Nat Rev Genet. 2008; 9:465–76. [PubMed: 18463664]
6. Schubeler D. Function and information content of DNA methylation. Nature. 2015; 517:321–6. [PubMed: 25592537]
7. Bestor TH. The DNA methyltransferases of mammals. Hum Mol Genet. 2000; 9:2395–402. [PubMed: 11005794]
8. Bird A. DNA methylation patterns and epigenetic memory. Genes Dev. 2002; 16:6–21. [PubMed: 11782440]
9. Reik W, Dean W, Walter J. Epigenetic reprogramming in mammalian development. Science. 2001; 293:1089–93. [PubMed: 11498579]
10. Wu SC, Zhang Y. Active DNA demethylation: many roads lead to Rome. Nat Rev Mol Cell Biol. 2010; 11:607–20. [PubMed: 20683471]
11. Smith ZD, Meissner A. DNA methylation: roles in mammalian development. Nature reviews Genetics. 2013; 14:204–20.
12. Jones P. Functions of DNA methylation: islands, start sites, gene bodies and beyond. Nature reviews Genetics. 2012; 13:484–492.
13. Wu H, Zhang Y. Reversing DNA methylation: mechanisms, genomics, and biological functions. Cell. 2014; 156:45–68. [PubMed: 24439369]
14. Pastor WA, Aravind L, Rao A. TETonic shift: biological roles of TET proteins in DNA demethylation and transcription. Nat Rev Mol Cell Biol. 2013; 14:341–56. [PubMed: 23698584]
15. Kriaucionis S, Heintz N. The nuclear DNA base 5-hydroxymethylcytosine is present in Purkinje neurons and the brain. Science. 2009; 324:929–30. [PubMed: 19372393]
16. Tahiliani M, et al. Conversion of 5-methylcytosine to 5-hydroxymethylcytosine in mammalian DNA by MLL partner TET1. Science. 2009; 324:930–5. [PubMed: 19372391]
17. Ito S, et al. Role of Tet proteins in 5mC to 5hmC conversion, ES-cell self-renewal and inner cell mass specification. Nature. 2010; 466:1129–33. [PubMed: 20639862]
18. He YF, et al. Tet-mediated formation of 5-carboxylcytosine and its excision by TDG in mammalian DNA. Science. 2011; 333:1303–7. [PubMed: 21817016]
19. Ito S, et al. Tet proteins can convert 5-methylcytosine to 5-formylcytosine and 5-carboxylcytosine. Science. 2011; 333:1300–3. [PubMed: 21778364]

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

20. Maiti A, Drohat AC. Thymine DNA glycosylase can rapidly excise 5-formylcytosine and 5-carboxylcytosine: Potential implications for active demethylation of CpG sites. J Biol Chem. 2011
21. Kohli RM, Zhang Y. TET enzymes, TDG and the dynamics of DNA demethylation. Nature. 2013; 502:472–9. [PubMed: 24153300]
22. Spruijt CG, et al. Dynamic readers for 5-(hydroxy)methylcytosine and its oxidized derivatives. Cell. 2013; 152:1146–59. [PubMed: 23434322]
23. Iurlaro M, et al. A screen for hydroxymethylcytosine and formylcytosine binding proteins suggests functions in transcription and chromatin regulation. Genome Biol. 2013; 14:R119. [PubMed: 24156278]
24. Kellinger MW, et al. 5-formylcytosine and 5-carboxylcytosine reduce the rate and substrate specificity of RNA polymerase II transcription. Nat Struct Mol Biol. 2012; 19:831–3. [PubMed: 22820989]
25. Wang L, et al. Molecular basis for 5-carboxycytosine recognition by RNA polymerase II elongation complex. Nature. 2015
26. Raiber EA, et al. 5-Formylcytosine alters the structure of the DNA double helix. Nat Struct Mol Biol. 2015; 22:44–9. [PubMed: 25504322]
27. Szulik MW, et al. Differential stabilities and sequence-dependent base pair opening dynamics of Watson-Crick base pairs with 5-hydroxymethylcytosine, 5-formylcytosine, or 5-carboxylcytosine. Biochemistry. 2015; 54:1294–305. [PubMed: 25632825]
28. Song CX, He C. Potential functional roles of DNA demethylation intermediates. Trends Biochem Sci. 2013
29. Ko M, et al. Impaired hydroxylation of 5-methylcytosine in myeloid cancers with mutant TET2. Nature. 2010; 468:839–43. [PubMed: 21057493]
30. Bachman M, et al. 5-Hydroxymethylcytosine is a predominantly stable DNA modification. Nat Chem. 2014; 6:1049–55. [PubMed: 25411882]
31. Shen L, et al. Genome-wide analysis reveals TET- and TDG-dependent 5-methylcytosine oxidation dynamics. Cell. 2013; 153:692–706. [PubMed: 23602152]
32. Song CX, et al. Genome-wide Profiling of 5-Formylcytosine Reveals Its Roles in Epigenetic Priming. Cell. 2013; 153:678–91. [PubMed: 23602153]
33. Inoue A, Shen L, Dai Q, He C, Zhang Y. Generation and replication-dependent dilution of 5fC and 5caC during mouse preimplantation development. Cell Res. 2011; 21:1670–6. [PubMed: 22124233]
34. Bachman M, et al. 5-Formylcytosine can be a stable DNA modification in mammals. Nat Chem Biol. 2015
35. Song CX, Yi C, He C. Mapping recently identified nucleotide variants in the genome and transcriptome. Nat Biotechnol. 2012; 30:1107–16. [PubMed: 23138310]
36. Plongthongkum N, Diep DH, Zhang K. Advances in the profiling of DNA modifications: cytosine methylation and beyond. Nat Rev Genet. 2014; 15:647–61. [PubMed: 25159599]
37. Booth MJ, Raiber EA, Balasubramanian S. Chemical methods for decoding cytosine modifications in DNA. Chem Rev. 2015; 115:2240–54. [PubMed: 25094039]
38. Booth MJ, et al. Quantitative sequencing of 5-methylcytosine and 5-hydroxymethylcytosine at single-base resolution. Science. 2012; 336:934–7. [PubMed: 22539555]
39. Wu H, et al. Genome-wide analysis of 5-hydroxymethylcytosine distribution reveals its dual function in transcriptional regulation in mouse embryonic stem cells. Genes Dev. 2011; 25:679–84. [PubMed: 21460036]
40. Pastor WA, et al. Genome-wide mapping of 5-hydroxymethylcytosine in embryonic stem cells. Nature. 2011; 473:394–7. [PubMed: 21552279]
41. Yu M, et al. Base-resolution analysis of 5-hydroxymethylcytosine in the mammalian genome. Cell. 2012; 149:1368–80. [PubMed: 22608086]
42. Stroud H, Feng S, Morey Kinney S, Pradhan S, Jacobsen SE. 5-Hydroxymethylcytosine is associated with enhancers and gene bodies in human embryonic stem cells. Genome Biol. 2011; 12:R54. [PubMed: 21689397]

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

43. Szulwach KE, et al. Integrating 5-hydroxymethylcytosine into the epigenomic landscape of human embryonic stem cells. PLoS Genet. 2011; 7:e1002154. [PubMed: 21731508]
44. Szulwach KE, et al. 5-hmC-mediated epigenetic dynamics during postnatal neurodevelopment and aging. Nat Neurosci. 2011; 14:1607–16. [PubMed: 22037496]
45. Lister R, et al. Global epigenomic reconfiguration during Mammalian brain development. Science. 2013; 341:1237905. [PubMed: 23828890]
46. Wang L, et al. Programming and inheritance of parental DNA methylomes in mammals. Cell. 2014; 157:979–91. [PubMed: 24813617]
47. Shen L, et al. Tet3 and DNA replication mediate demethylation of both the maternal and paternal genomes in mouse zygotes. Cell Stem Cell. 2014; 15:459–70. [PubMed: 25280220]
48. Guo F, et al. Active and passive demethylation of male and female pronuclear DNA in the Mammalian zygote. Cell Stem Cell. 2014; 15:447–58. [PubMed: 25220291]
49. Sun Z, et al. A sensitive approach to map genome-wide 5-hydroxymethylcytosine and 5-formylcytosine at single-base resolution. Mol Cell. 2015; 57:750–61. [PubMed: 25639471]
50. Song CX, et al. Selective chemical labeling reveals the genome-wide distribution of 5-hydroxymethylcytosine. Nat Biotechnol. 2011; 29:68–72. [PubMed: 21151123]
51. Booth MJ, Marsico G, Bachman M, Beraldi D, Balasubramanian S. Quantitative sequencing of 5-formylcytosine in DNA at single-base resolution. Nature Chemistry. 2014
52. Lu X, et al. Chemical Modification-Assisted Bisulfite Sequencing (CAB-Seq) for 5-Carboxylcytosine Detection in DNA. J Am Chem Soc. 2013; 135:9315–7. [PubMed: 23758547]
53. Lu X, et al. Base-resolution maps of 5-formylcytosine and 5-carboxylcytosine reveal genome-wide DNA demethylation dynamics. Cell Res. 2015; 25:386–9. [PubMed: 25591929]
54. Wu H, Wu X, Shen L, Zhang Y. Single-base resolution analysis of active DNA demethylation using methylase-assisted bisulfite sequencing. Nat Biotechnol. 2014; 32:1231–40. [PubMed: 25362244]
55. Neri F, et al. Single-Base Resolution Analysis of 5-Formyl and 5-Carboxyl Cytosine Reveals Promoter DNA Methylation Dynamics. Cell Rep. 2015
56. Ooi SK, et al. DNMT3L connects unmethylated lysine 4 of histone H3 to de novo methylation of DNA. Nature. 2007; 448:714–7. [PubMed: 17687327]
57. Boulard M, Edwards JR, Bestor TH. FBXL10 protects Polycomb-bound genes from hypermethylation. Nat Genet. 2015

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

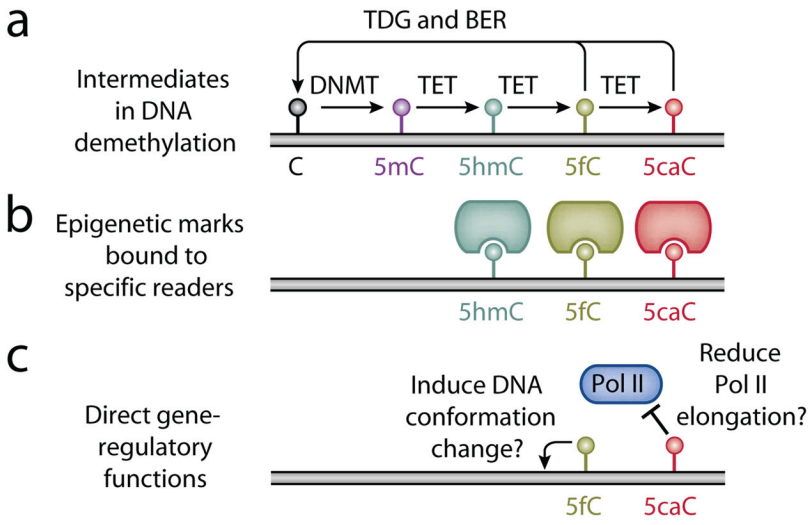

Figure 1. Schematic diagram of potential functions for 5hmC, 5fC and 5caC

(a) Oxidized methylcytosines (i.e. 5hmC/5fC/5caC) serve as intermediate products in TET/TDG-mediated active DNA demethylation process.

(b) All oxidized cytosine bases may act as stable or transient epigenetic marks by attracting or repelling specific DNA binding proteins.

(c) 5fC and 5caC may have additional gene regulatory functions, including retarding RNA polymerase II elongation or altering DNA conformation.

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

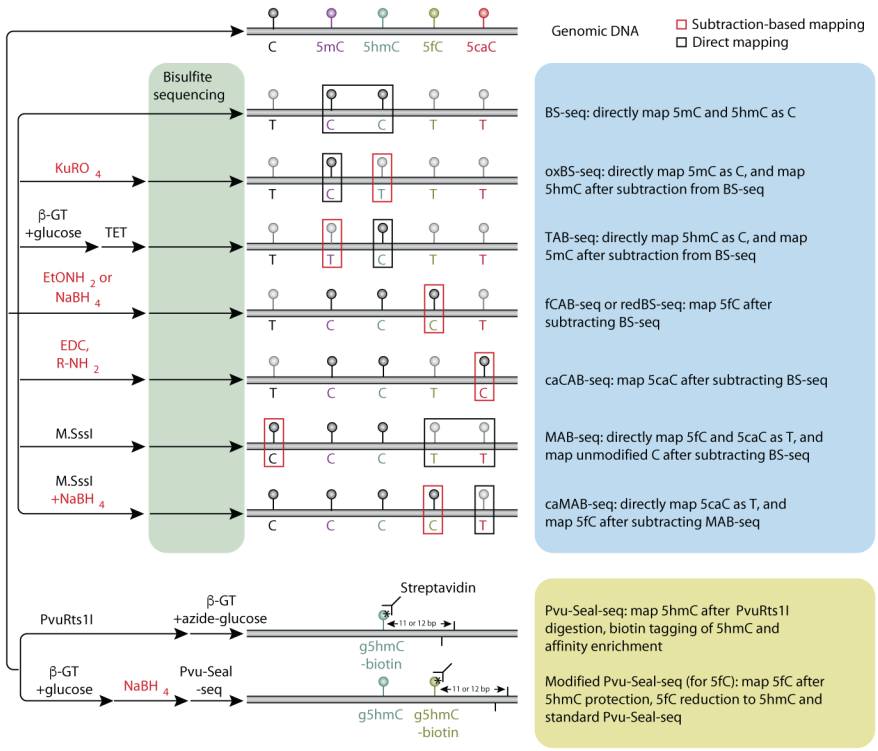

Figure 2. Schematic diagram of base-resolution mapping methods for 5hmC, 5fC and 5caC

DNMT methylates unmodified C to generate 5mC, which can be successively oxidized by TET to generate 5hmC/5fC/5caC. Highly oxidized cytosine variatnts, 5fC and 5caC, are repared by TDG/BER to regenerate unmodified C. In conjunction with various chemical (in red) and enzymatic (in black) treatment, bisulfite sequencing (BS-seq) or 5hmC-dependent endonuclease PvuRts1l-based method have been developed to map unmodified C ("MAB-seq sub BS-seq"), 5mC (oxBS-seq, "BS-seq sub TAB-seq"), 5hmC ("BS-seq sub oxBS-seq", TAB-Seq and Pvu-Seal-seq), 5fC ("fCAB-Seq sub BS-seq", "redBS-seq sub BS-seq", "caMAB-seq sub MAB-seq" and modified Pvu-Seal-seq), 5caC ("caCAB-Seq sub BS-seq" and caMAB-seq), 5fC/5caC (MAB-seq). In contrast to direct mapping methods (e.g. oxBS-seq for 5mC, TAB-seq for 5hmC and MAB-seq for 5fC/5caC), other BS-seq-based mapping strategies (e.g. oxBS-seq for 5hmC, fCAB-seq for 5fC and caCAB-seq for 5caC) require subtracting signals between conventional BS-Seq from those of modified BS-Seq to indirectly determine the position and abundance of oxidized 5mC bases. In Pvu-Seal-seq, genomic DNA is first digested with PvuRts1l. This endonuclease recognizes 5hmC and creates a double-stranded break 11–12 bp downstream, leaving a 2 bp 3′ overhang. While PvuRts1l exhibits highest activity on 5hmC, it also recognizes 5mC or C, albeti with a lower cleavage activity. To increase the specificity of 5hmC mapping, a chemical labeling step (tag 5hmC with biotin) is needed to selectively enrich 5hmC-containing DNA fragments after PvuRts1l digestion. Unlike BS-seq-based mapping methods, Pvu-Seal-seq involves a affinity enrichment step and cannot quantify the absolute levels of 5hmC or 5fC sites.

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

Table 1

Comparison of different base-resolution mapping methods for oxidized methylcytosines

	Advantage	Disadvantage	No. of oxidized 5mCs in mESCs (mass spectrometry)	No. of oxidized 5mC in mESCs (genomic sequencing)	References
5hmC			6.25 million (1,250 p.p.m., assuming ~20% per site)		
oxBS-seq	No enzymatic treatment	Subtraction based; needs deep sequencing		NA	38
TAB-seq	Direct mapping of 5hmC	Potential incomplete enzymatic treatment		2.1 million	41,46
Pvu-Seal-seq	Higher sensitivity	Cannot determine absolute 5hmC level		20.8 million	49
Third-generation sequencer	No need for PCR	Need to increase accuracy and throughput		NA	36,37
5fC			0.19 million (wild type, 19 p.p.m., assuming ~10% per site) or 0.95 million (Tdg mutant; ~5-fold increase compared to wild type)		
fCAB-seq	No enzymatic treatment	Subtraction based; needs deep sequencing		NA	32,46,53
redBS-seq	No enzymatic treatment	Subtraction based; needs deep sequencing		NA	51
MAB-seq	Direct and simultaneous map - ping of 5fC and 5caC	Cannot determine 5fC and 5caC in non-CpG		0.3 million (wild type) or 0.7 million (Tdg mutant)	48,54,55
Pvu-Seal-seq (modified for 5fC)	Higher sensitivity	Cannot determine absolute 5fC level		6.2 million (wild type)	49
Third-generation sequencer	No need for PCR	Need to increase accuracy and throughput		NA	36,37
5caC			0.034 million (wild type, 3.4 p.p.m., assuming ~10% per site) or 0.17 million (Tdg mutant; ~5-fold increase compared to wild type)		
caCAB-seq	No enzymatic treatment	Subtraction based; needs deep sequencing		NA	52,53
caMAB-seq	Direct mapping of 5caC	Cannot determine 5caC in non-CpG		NA	54
Third-generation sequencer	No need for PCR	Need to increase accuracy and throughput		NA	36,37

NA, not applicable.

Nat Struct Mol Biol. Author manuscript; available in PMC 2015 December 07.