nature structural & molecular biology

Charting oxidized methylcytosines at base resolution

Hao Wu[1–4] & Yi Zhang[1–4]

DNA cytosine methylation is a key epigenetic mark that is required for normal mammalian development. Iterative oxidation of 5-methylcytosine (5mC) by the TET family of DNA dioxygenases generates three oxidized nucleotides: 5-hydroxymethylcytosine (5hmC), 5-formylcytosine (5fC) and 5-carboxylcytosine (5caC). Recent advances in genomic mapping techniques have suggested that these oxidized cytosines not only function in the process of active reversal of 5mC but also may possess unique regulatory functions in the mammalian genome.

Methylation of DNA allows the genome to carry regulatory information beyond its canonical role as a genetic blueprint. In bacteria, methylation on either adenosine or cytosine affects diverse biological processes such as degrading foreign DNA, tracking mismatch repair and regulating DNA replication[1]. Methylation at the 5 position of cytosine (forming 5mC) is evolutionarily conserved in many eukaryotic organisms and is functionally linked to regulation of gene expression and maintenance of genome integrity[2]. Intriguingly, recent studies have indicated that adenosine methylation (forming, e.g., N^6-methyladenosine) occurs in *Caenorhabditis elegans* and *Drosophila melanogaster*, two organisms that were previously thought to lack 5mC[3,4], thus raising the possibility that DNA methylation has a general role in eukaryotic biology.

In vertebrates, 5mC is the predominant DNA modification, and it occurs throughout the entire genome, thus suggesting that methylation might be a default state[5,6]. The *de novo* DNA methyltransferases, DNMT3A and DNMT3B, primarily catalyze the formation of 5mC at palindromic CpG dinucleotides, and the maintenance DNA methyltransferase, DNMT1, enables faithful propagation of CpG methylation patterns through cell divisions[7]. Heritable CpG methylation (forming mCpG) is therefore considered to be a classic epigenetic mark involved in long-term epigenetic memory processes such as genomic imprinting, X-chromosome inactivation and silencing of repeats[8]. Interestingly, highly dynamic changes in DNA methylation occur throughout the genome during early embryonic development and are required for critical biological processes such as erasure of parental origin–specific imprints in developing primordial germ cells[9,10]. In addition, genome-wide mapping of 5mC has revealed that active gene-regulatory sequences, such as 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 activities may exist for active 5mC removal in vertebrates. The recent identification of the ten-eleven translocation (TET) family of 5mC dioxygenases has provided a plausible pathway to catalyze active DNA demethylation[13,14]. TET proteins convert 5mC into 5-hydroxymethylcytosine (5hmC)[15–17]. Further successive oxidations mediated by TET result in formation of 5-formylcytosine (5fC) and 5-carboxylcytosine (5caC)[18,19], both of which are efficiently excised by thymine DNA glycosylase (TDG) and restored to unmodified cytosines through the base excision repair (BER) pathway[18,20,21]. Genetic studies of TET-mutant mice have indicated that these 5mC oxidases have important roles in embryonic development, stem-cell differentiation, genomic imprinting, neuronal functions and cancer (reviewed in refs. 13,14).

Growing evidence has indicated that oxidized methylcytosines, in addition to having roles as intermediates of active DNA demethylation (**Fig. 1a**), may also possess unique regulatory functions. Screens for 'reader' proteins of 5hmC, 5fC and 5caC[22,23] (**Fig. 1b**) have identified not only candidates linked to DNA-repair processes but also transcription factors and chromatin-modifying enzymes. Interestingly, the number of identified candidates for 5fC and 5caC is much higher than for 5hmC, thus possibly reflecting the unique chemical properties of the formyl and carboxyl groups of these two highly oxidized bases. Biochemical and structural evidence has also indicated that 5fC and 5caC within gene bodies may decrease the elongation rate of RNA polymerase II[24,25] (**Fig. 1c**). Furthermore, biophysical studies have revealed that these oxidized bases may influence base-pairing and DNA structure[26,27] and may thus affect DNA-templated processes through direct effects on DNA conformation. Finally, single- or double-strand breaks associated with the DNA-repair process downstream of excision of 5fC and 5caC may contribute to gene regulation as well[13].

Understanding the mechanisms underlying these roles requires the ability to comprehensively profile the distribution of oxidized cytosines in the mammalian genome. Recent technological advances have produced genomic maps of oxidized 5mC bases at unprecedented resolution, revealing that TET-mediated 5mC-oxidation events are preferentially targeted to genomic regions associated with regulatory functions. Despite these intriguing correlations, exactly how oxidized 5mC bases exert their function at these regulatory regions is largely unclear and under active investigation[28]. Here we summarize recent advances in genomic mapping methods for 5hmC, 5fC and 5caC modified bases, and highlight the potential functions of oxidized 5mC derivatives in gene regulation.

[1]Howard Hughes Medical Institute, Boston, Massachusetts, USA. [2]Program in Cellular and Molecular Medicine, Boston Children's Hospital, Boston, Massachusetts, USA. [3]Department of Genetics, Harvard Medical School, Boston, Massachusetts, USA. [4]Harvard Stem Cell Institute, Harvard Medical School, Boston, Massachusetts, USA. Correspondence should be addressed to Y.Z. (yzhang@genetics.med.harvard.edu).

Received 15 May; accepted 16 July; published online 3 September 2015; doi:10.1038/nsmb.3071

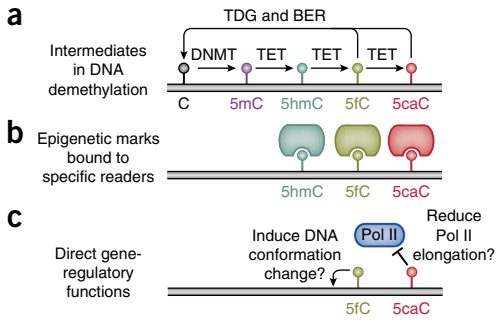

Figure 1 Schematic diagram of potential functions for 5hmC, 5fC and 5caC. (**a**) Oxidized methylcytosines 5hmC, 5fC and 5caC serve as intermediate products in TET- and TDG-mediated active DNA-demethylation processes. (**b**) All oxidized cytosine bases may act as stable or transient epigenetic marks by attracting or repelling specific DNA-binding proteins. (**c**) 5fC and 5caC may have additional gene-regulatory functions, including retarding RNA polymerase II (Pol II) elongation and altering DNA conformation.

Genomic mapping of oxidized 5mC 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 (~4% of total cytosines) that are stable across somatic tissues, 5hmC levels exhibit tissue-specific variation; 5hmC can be as high as 40% of the 5mC level in Purkinje neurons in the cerebellum[15], ~5% of 5mC in mouse embryonic stem cells (ESCs)[16,17] and as low as ~1% of 5mC in some immune cells[29]. Although 5hmC has been considered to be an intermediate of TET-mediated DNA demethylation, recent studies have suggested that the majority of 5hmC modifications are stable in mouse tissues[30]. Iterative oxidation of 5hmC by TET generates 5fC and 5caC, which are present at ~100-fold-lower levels than 5hmC in wild-type ESCs[19]. The low abundance of 5fC (~20 parts per million (p.p.m.) cytosines in ESCs) and 5caC (~3 p.p.m.) is due in part to the robust excision activity of 5fC and 5caC by TDG in mammalian cells. Indeed, depletion of TDG in mouse ESCs (mESCs) results in a 5- to 10-fold increase in 5fC and 5caC levels[31,32], thus suggesting that the majority of 5fC and 5caC is only transiently present in the mammalian genome. One potential exception is the initial generation and eventual replication-dependent dilution of 5fC and 5caC in early preimplantation embryos[33], in which TDG mRNAs are not detected. Using a stable-isotope tracing strategy, a recent study has indicated that 5fC can also be a stable modification in nonproliferating cells or postnatal tissues[34], albeit at very low levels (ranging from 0.2 to 15 p.p.m.). The global level of oxidized cytosine bases can be quantified by several methods, including thin-layer chromatography, modification-specific antibodies, chemical tagging and mass spectrometry (reviewed in ref. 35).

Genome-wide mapping of oxidized 5mC variants is technically challenging, owing to their extremely low abundance in mammalian cells. Early mapping efforts relied primarily on affinity-enrichment methods that used either modification-specific antibodies or chemical tagging[36,37]. However, the genomic maps generated by these methods are of limited resolution, reveal only relative enrichment over control assays and lack precise strand-distribution information. Recent new methods that distinguish 5mC-oxidation variants at single-base resolution and genomic scale have provided important new insights into the biological functions of these oxidized cytosine bases.

Base-resolution mapping of 5hmC. The current method of choice for single-base-resolution mapping of cytosine methylation 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 by bisulfite treatment and are consequently read out as cytosine (C) after PCR amplification, whereas unmodified C, 5fC and 5caC are deaminated 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 methods have been developed (**Fig. 2**).

The first, termed oxidative bisulfite sequencing (oxBS-seq), uses potassium perruthenate ($KRuO_4$) to specifically oxidize 5hmC to 5fC. Thus, 5hmC, which is converted to 5fC, is detected as T, and 5mC remains intact and is read out as C. Whereas 5mC is mapped directly, the absolute level and precise location of 5hmC are determined by subtracting signals of oxBS-seq maps (5mC) from those of traditional BS-seq (5mC and 5hmC)[38]. Because this involves subtraction between two random sampling–based BS-seq experiments, very deep sequencing coverage is necessary to achieve high-confidence 5hmC mapping. One strategy to decrease the sequencing effort needed combines oxBS-seq with a reduced representation BS-seq (RRBS) strategy[38]. In this oxRRBS approach, converted DNA is first digested with the enzyme MspI (which recognizes C^CGG, with ^ indicating the cleavage site) to enrich for CpG-containing DNA fragments, thus permitting deep (~120× per cytosine) and selective sequencing of genomic CpG-rich sequences such as CpG islands and repeat elements. oxRRBS analysis of mESCs has revealed that 5hmC is relatively enriched at transcriptionally poised CpG-rich promoters, which drive expression of lineage-specific transcription factors during cellular differentiation. Indeed, 5hmC enrichment at promoters is negatively correlated with steady-state transcription levels of the genes that they control, an observation consistent with 5hmC maps generated by affinity enrichment–based methods[39,40].

The second base-resolution 5hmC-mapping method, called TET-assisted bisulfite sequencing (TAB-seq), involves TET-mediated enzymatic conversion of non-5hmC modifications (5mC and 5fC) to 5caC and subsequent bisulfite sequencing[41]. In TAB-seq, 5hmC is first protected from TET-mediated oxidation via glucosylation by β-glucosyltransferase, and 5mC and 5fC are subsequently oxidized to 5caC by recombinant TET1 protein at high concentration. Thus, C, 5mC, 5fC and 5caC are detected as T, and only 5hmC is read as C. Here, the advantage is that 5hmC is mapped directly, and this decreases the sequencing effort needed for high-confidence mapping[41]. With a medium sequencing depth (17× to 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 million and ~2.1 million high-confidence 5hmC sites have been identified in human and mouse ESCs, respectively, and have been mapped at base resolution (**Table 1**). 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 data sets. Previous affinity enrichment–based maps have also suggested that 5hmC is enriched at distal *cis*-regulatory elements[39,42,43], but these lacked the resolution required to determine the relationship between 5hmC position and transcription factor (TF)-binding motifs. The base-resolution map revealed that 5hmC is typically not located within TF-binding sites but instead is enriched in regions immediately adjacent to TF motifs, thus generating bimodal peaks of 5hmC flanking the motif. The nervous system possesses the highest level of 5hmC among mammalian somatic tissues, and age-dependent increases of total 5hmC levels

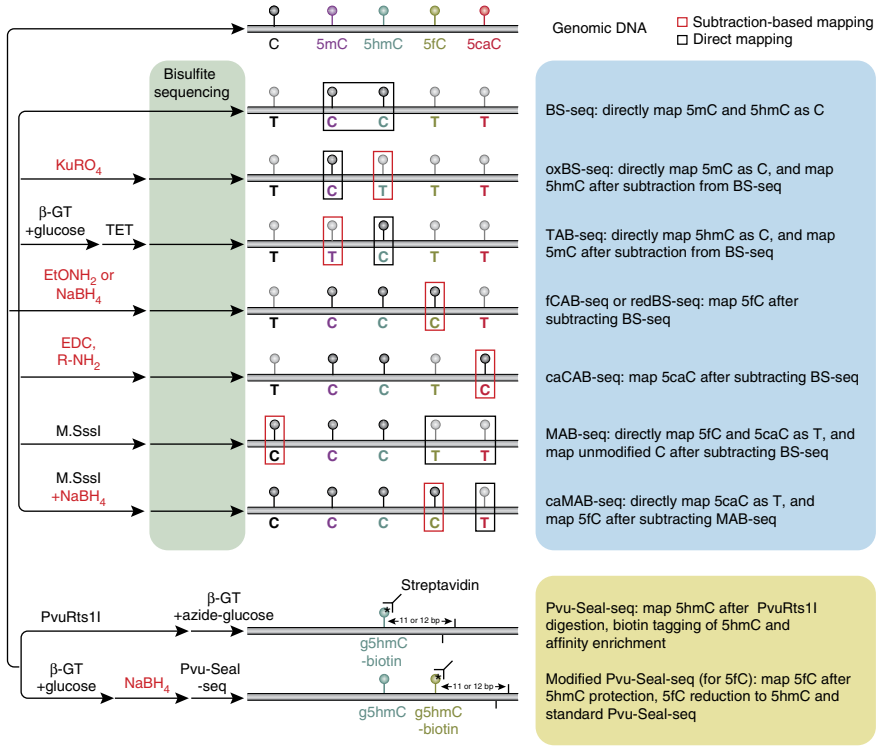

Figure 2 Schematic diagram of base-resolution mapping methods for 5hmC, 5fC and 5caC. DNMT methylates unmodified C to generate 5mC, which can be successively oxidized by TET to generate 5hmC, 5fC and 5caC. Highly oxidized cytosine variants, 5fC and 5caC, are repaired by TDG and BER to regenerate unmodified C. In conjunction with various chemical (in red) and enzymatic (in black) treatments, bisulfite sequencing (BS-seq) and methods based on the 5hmC-dependent endonuclease PvuRts1I have been developed to map unmodified C (MAB-seq sub BS-seq), 5mC (oxBS-seq and 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) and both 5fC and 5caC together (MAB-seq). In contrast to direct mapping methods (for example, oxBS-seq for 5mC, TAB-seq for 5hmC and MAB-seq for 5fC and 5caC), other BS-seq–based mapping strategies (for example, oxBS-seq for 5hmC, fCAB-seq for 5fC and caCAB-seq for 5caC) require subtracting signals of 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 PvuRts1I. This endonuclease recognizes 5hmC and creates a double-stranded break 11 or 12 bp downstream, leaving a 2 bp 3′ overhang. Although PvuRts1I exhibits its highest activity on 5hmC, it also recognizes 5mC or C but has a lower cleavage activity toward them. 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 PvuRts1I digestion. Unlike BS-seq–based mapping methods, Pvu-Seal-seq involves an affinity-enrichment step and cannot quantify the absolute levels of 5hmC or 5fC sites. β-GT, β-glucosyltransferase.

in the cortex, hippocampus and cerebellum support a role of 5hmC as a stable epigenetic mark in neuronal genomes[44]. Indeed, whole-genome TAB-seq analysis of mammalian fetal and young-adult frontal cortexes not only has confirmed a general increase of 5hmC at various genomic loci in adult compared to fetal brains but also has revealed a positive correlation between intragenic 5hmC enrichment in a CpG context and transcriptional activity[45]. TAB-seq has also been used to analyze the 5hmC pattern in two-cell embryos[46]. When combined with allele-specific analysis, TAB-seq at a sequencing depth of 18× per cytosine identified ~0.1 million 5hmCpGs in the paternal genome and ~0.12 million 5hmCpGs in the maternal genome, thus suggesting that TET-mediated 5mC oxidation takes place on both sperm- and oocyte-derived chromosomes. Indeed, genetic analysis of wild-type and Tet3 maternal knockout (KO) embryos has indicated that Tet3 deficiency affects DNA demethylation of both paternal and maternal genomes in one-cell zygotes[47,48].

Owing to low levels of genomic 5hmC, base-resolution mapping methods such as oxBS-seq and TAB-seq require ultra-deep sequencing coverage (>100×) to achieve high-confidence identification of low-abundance 5hmC sites (<5% modification level). To overcome this limitation, a more sensitive approach was developed that relies on the PvuRst1l family of 5hmC-dependent restriction endonucleases[49]. PvuRts1l enzymes create double-stranded breaks 11 or 12 bp downstream of 5hmC, generating DNA fragments with 2-bp 3′ overhangs that can then be converted into sequencing libraries. The location of 5hmC can then be determined by mapping the cleavage sites. Combining a chemical labeling–based 5hmC enrichment method, hMe-Seal[50], with PvuRst1l digestion allows cost-efficient genome-wide 5hmC mapping at base resolution. This 'Pvu-Seal-seq' approach has identified 20.8 million 5hmC sites in two technical replicates of mESCs (**Table 1**), a number of sites ten times higher than that observed by whole-genome TAB-seq analysis. 24% of the 5hmC sites that have been identified by Pvu-Seal-seq are in non-CpG contexts (~50% in CpA), a surprising result given that both oxRRBS and whole-genome TAB-seq analyses have suggested that nearly all 5hmC sites (~99%) are detected in the CpG context[38,41]. Interestingly, Pvu-Seal-seq analysis of proliferating mESCs indicated that 64% of 5hmCpG sites were conserved between two biological replicates, but only 24% of 5hmC in non-CpG contexts (5hmCpH) were reproducibly detected[49]. These results imply that 5hmCpH is much less stable than 5hmCpG, possibly because 5mC in non-CpG contexts is not faithfully maintained during cell division. Locus-specific TAB-seq analysis (sequencing depth >100×) confirmed some of the newly identified 5hmCpH sites and revealed that the average 5hmC level in non-CpG contexts (2.8%) was significantly lower than that of 5hmCpG (11.4%), highlighting the need of ultra-high sequencing depth to detect 5hmCpH with high confidence. Overall, these results show that Pvu-Seal-seq has the enhanced sensitivity required to detect low-abundance and/or unstable 5hmC sites in the genome without the need for ultra-high sequencing depth. However, owing to the initial affinity-enrichment step, Pvu-Seal-seq cannot determine the absolute percentage of 5hmC and thus can be used only to quantify relative changes in 5hmC levels.

Base-resolution mapping of 5fC and 5caC. Available evidence has suggested 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 excision of 5fC and 5caC and subsequent BER[13]. The active

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 mapping 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.

DNA-demethylation pathway involving generation and excision repair of 5fC and 5caC is of particular interest because it occurs in both proliferating and postmitotic cells. The observation that 5hmCpG is relatively stable *in vivo* suggests that identifying methylated CpGs that are targeted for active DNA demethylation requires the ability to quantify 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 to map these oxidized cytosines at single-base resolution (**Fig. 2**).

Subtraction-dependent methods. This group of mapping techniques uses specific chemical treatment to protect either 5fC or 5caC from deamination by bisulfite treatment. Thus, 5fC chemically assisted bisulfite sequencing (fCAB-seq) uses *O*-ethylhydroxylamine ($EtONH_2$) to protect 5fC[32], whereas a similar approach called reduced bisulfite sequencing (redBS-seq) uses sodium borohydride ($NaBH_4$) to selectively reduce 5fC to 5hmC[51]. To protect 5caC, 5caC chemically assisted bisulfite sequencing (caCAB-seq) 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 and 5caC, coupled with the possibility that chemicals used in these methods may react with DNA moieties other than their intended targets (e.g., reaction of $EtONH_2$ with abasic sites or unmodified C) may complicate the interpretation of results. To decrease the sequencing effort needed, various enrichment strategies have been integrated with subtraction-dependent 5fC- and 5caC-mapping methods to generate genome-scale maps. By combining chromatin immunoprecipitation (using antibodies against monomethylated Lys4 on histone H3 (H3K4me1)) with fCAB-seq (termed H3K4me1-fCAB-seq), 5fC abundance can be examined within the fraction of the genome associated with active or poised enhancer activity[32]. Integrating redBS-seq and RRBS allows for 5fC mapping within genomic regions containing CpG-rich sequences (mostly gene promoters)[51]. Importantly, the MspI enzyme used in standard RRBS only partially digests 5fC-containing C^CGG sites and completely fails to cut 5caCpGs[19]. Thus, either reducing 5fC to 5hmC before MspI digestion or choosing an enzyme (for example, TaqαI, T^CGA) that is less influenced by 5fC and 5caC modification should be considered when selecting the RRBS strategy for analysis of 5fC and 5caC. More recently, DNA immunoprecipitation with antibodies specific to 5fC or 5caC has been used to enrich DNA fragments containing these rare modified bases before fCAB-seq and caCAB-seq assays. Although this affinity enrichment–based strategy can precisely map the positions of 5fC and 5caC, only relative enrichment of 5fC or 5caC can be determined[53]. Thus, subtraction-based methods for mapping of 5fC and 5caC are not well suited for genome-scale quantitative analysis of 5fC and 5caC at base resolution.

Subtraction-independent methods. Because 5fC and 5caC are present at extremely low levels in the genome, it is highly desirable to directly map these rare bases. Moreover, because both 5fC and 5caC are

substrates for TDG-mediated excision repair, determining the strand preference of TET- and TDG-mediated demethylation requires both 5fC and 5caC to be simultaneously mapped in a single experiment. To circumvent these limitations, three groups independently developed a modified BS-seq method, called methylase-assisted bisulfite sequencing (MAB-seq), enabling direct and simultaneous mapping of 5fC and 5caC at single-base resolution[48,54,55]. In MAB-seq, genomic DNA is first treated with the bacterial DNA CpG methyltransferase M.SssI, an enzyme that efficiently methylates CpG dinucleotides. Bisulfite conversion of methylase-treated DNA leads to deamination of 5fC and 5caC only; because CpGs that were unmodified in the original samples are protected as 5mCpG, subsequent sequencing can directly reveal 5fC and 5caC as T, whereas C, 5mC and 5hmC are read out as C. Notably, MAB-seq is unable to distinguish 5fC and 5caC from unmodified C within a non-CpG context, owing to the poor activity of M.SssI toward CpH.

Our own group has further developed a method, 5caC methylase–assisted bisulfite sequencing (caMAB-seq)[54], to directly map 5caC at single-base resolution. This modified version of MAB-seq takes advantage of the ability of $NaBH_4$ to selectively reduce 5fC to 5hmC. When M.SssI-treated DNA is subsequently incubated with $NaBH_4$, only 5caC is read out as T; the original 5fC, along with C, 5mC and 5hmC, is read as C. A modified Pvu-Seal-seq method integrating $NaBH_4$-mediated 5fC reduction into the Pvu-Seal-seq workflow can also directly detect 5fC at base resolution.

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

A whole-genome base-resolution map of 5fC and 5caC has recently been generated by MAB-seq for wild-type and *Tdg*-depleted mESCs[54,55]. Deep-sequencing analysis (>50× per cytosine) of wild-type ESCs has identified ~0.3 million 5fC- and 5caC-modified CpGs (5fC and 5caCpGs) with an average modification level of 8–10% (ref. 55 and **Table 1**). Interestingly, 5fC and 5caCpGs are enriched at active gene promoters and enhancers in wild-type ESCs, colocalizing with both TET and TDG proteins[55]. A reduced-representation version of MAB-seq (RRMAB-seq) that allows more focused analysis of 5fC and 5caC at CpG-rich promoters in wild-type and Tdg-depleted cells has revealed a dynamic DNA methylation-and-demethylation process at actively transcribed gene promoters. Because both the increase of 5mC in *Tet1*- and *Tet2*-depleted cells (from ~2% to ~5%) and the increase of 5fC and 5caC in *Tdg*-depleted cells are quite modest, the TET- and TDG-mediated active DNA-demethylation pathway may have only a relatively minor role in protecting these CpG-rich promoters from hypermethylation. Redundant 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.

With use of *Dnmt1*, *Dnmt3a* and *Dnmt3b* triple-KO and *Tet1* and *Tet2* double-KO mESCs (which have 5fC- and 5caC-free genomes) as reference controls for false discovery rate, whole-genome base-resolution 5fC and 5caC mapping (14× per cytosine) in *Tdg*-depleted mESCs has identified ~0.7 million 5fC and 5caCpGs with an average modification level of 20–30% (ref. 54 and **Table 1**). 5fC and 5caCpGs are most enriched at DNase I–hypersensitive sites, H3K4me1-marked ESC enhancers, CTCF-bound insulators and exonic regions. Strand-specific comparative analysis of 5fC, 5caCpG and 5hmCpG mapping has shown that >90% of 5hmCpGs do not overlap with 5fC or 5caCpGs, thus suggesting that proteins exhibit distinct catalytic processivities at different CpGs. Further analysis has suggested that the processivity of TET-mediated iterative oxidation correlates with local chromatin accessibility, such that 5fC and 5caCpGs (including dual-modified CpGs), compared to 5hmC-alone CpGs, are associated with higher levels of DNase I–hypersensitive sites; histone variants H2A.Z and H3.3, which destabilize nucleosome structure; and pluripotency-related TFs (Oct4, Nanog and Sox2)[54]. Similarly to the observed strand asymmetry of 5hmCpGs[41], 5fC and 5caCpGs also exhibit a strong preference for asymmetrical modification[54,55], thus suggesting that single-strand breaks rather than double-strand breaks are the predominant intermediates during TET- and TDG-mediated active DNA-demethylation process. Although 5hmC is rare and/or unstable in the non-CpG context, the presence of a low abundance of 5fC and 5caCpHs cannot be excluded. Indeed, base-resolution 5fC mapping by Pvu-Seal-seq has identified 6.2 million 5fC sites (75% in CpG and 25% in CpH) in wild-type mESCs[49] (**Table 1**), a number significantly higher than the number of 5fC and 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 efforts to map and quantify these modifications in different cell and tissue types, particularly in ESCs and developing brain tissues. The major challenge that such efforts must overcome is the scarcity of 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 has suggested that 5hmC, 5fC and 5caC are not randomly distributed in the mammalian genome but rather are 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 notable new insights into the mechanism and function of TET- and TDG-mediated active DNA-demethylation processes. However, new experimental methods and strategies are needed to allow integrated analysis of all five distinct cytosine modification states in small numbers of cells, particularly in preimplantation embryos and primordial germ cells in which dynamic global DNA-methylation changes are observed. Emerging technologies such as SMRT and nanopore sequencing have the potential for direct reading of DNA modifications on single molecules without the need for DNA amplification or bisulfite conversion[36,37] (**Table 1**). Still, the throughput and accuracy of these third-generation sequencing methods need to be substantially improved before they can be used for mapping low-abundance cytosine modifications in complex mammalian genomes. We anticipate that future studies applying methods of single-base-resolution mapping to various biological and pathological systems may greatly advance understanding of the genomic functions of oxidized cytosine bases.

ACKNOWLEDGMENTS
We thank L.M. Tuesta for critical reading of the manuscript. This work was supported by US National Institutes of Health 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 US National Human Genome Research Institute (K99HG007982). Y.Z. is supported as an Investigator of the Howard Hughes Medical Institute.

COMPETING FINANCIAL INTERESTS
The authors declare no competing financial interests.

1. Arber, W. & Dussoix, D. Host specificity of DNA produced by *Escherichia coli*. I. Host controlled modification of bacteriophage lambda. *J. Mol. Biol.* **5**, 18–36 (1962).
2. Law, J.A. & Jacobsen, S.E. Establishing, maintaining and modifying DNA methylation patterns in plants and animals. *Nat. Rev. Genet.* **11**, 204–220 (2010).
3. Greer, E.L. *et al.* DNA methylation on N6-adenine in *C. elegans*. *Cell* **161**, 868–878 (2015).
4. Zhang, G. *et al.* N6-methyladenine DNA modification in *Drosophila*. *Cell* **161**, 893–906 (2015).
5. Suzuki, M.M. & Bird, A. DNA methylation landscapes: provocative insights from epigenomics. *Nat. Rev. Genet.* **9**, 465–476 (2008).
6. Schübeler, D. Function and information content of DNA methylation. *Nature* **517**, 321–326 (2015).
7. Bestor, T.H. The DNA methyltransferases of mammals. *Hum. Mol. Genet.* **9**, 2395–2402 (2000).
8. Bird, A. DNA methylation patterns and epigenetic memory. *Genes Dev.* **16**, 6–21 (2002).
9. Reik, W., Dean, W. & Walter, J. Epigenetic reprogramming in mammalian development. *Science* **293**, 1089–1093 (2001).
10. Wu, S.C. & Zhang, Y. Active DNA demethylation: many roads lead to Rome. *Nat. Rev. Mol. Cell Biol.* **11**, 607–620 (2010).
11. Smith, Z.D. & Meissner, A. DNA methylation: roles in mammalian development. *Nat. Rev. Genet.* **14**, 204–220 (2013).
12. Jones, P.A. Functions of DNA methylation: islands, start sites, gene bodies and beyond. *Nat. Rev. Genet.* **13**, 484–492 (2012).
13. Wu, H. & Zhang, Y. Reversing DNA methylation: mechanisms, genomics, and biological functions. *Cell* **156**, 45–68 (2014).
14. Pastor, W.A., Aravind, L. & Rao, A. TETonic shift: biological roles of TET proteins in DNA demethylation and transcription. *Nat. Rev. Mol. Cell Biol.* **14**, 341–356 (2013).
15. Kriaucionis, S. & Heintz, N. The nuclear DNA base 5-hydroxymethylcytosine is present in Purkinje neurons and the brain. *Science* **324**, 929–930 (2009).
16. Tahiliani, M. *et al.* Conversion of 5-methylcytosine to 5-hydroxymethylcytosine in mammalian DNA by MLL partner TET1. *Science* **324**, 930–935 (2009).
17. Ito, S. *et al.* Role of Tet proteins in 5mC to 5hmC conversion, ES-cell self-renewal and inner cell mass specification. *Nature* **466**, 1129–1133 (2010).
18. He, Y.F. *et al.* Tet-mediated formation of 5-carboxylcytosine and its excision by TDG in mammalian DNA. *Science* **333**, 1303–1307 (2011).
19. Ito, S. *et al.* Tet proteins can convert 5-methylcytosine to 5-formylcytosine and 5-carboxylcytosine. *Science* **333**, 1300–1303 (2011).
20. Maiti, A. & Drohat, A.C. Thymine DNA glycosylase can rapidly excise 5-formylcytosine and 5-carboxylcytosine: potential implications for active demethylation of CpG sites. *J. Biol. Chem.* **286**, 35334–35338 (2011).
21. Kohli, R.M. & Zhang, Y. TET enzymes, TDG and the dynamics of DNA demethylation. *Nature* **502**, 472–479 (2013).
22. Spruijt, C.G. *et al.* Dynamic readers for 5-(hydroxy)methylcytosine and its oxidized derivatives. *Cell* **152**, 1146–1159 (2013).
23. Iurlaro, M. *et al.* A screen for hydroxymethylcytosine and formylcytosine binding proteins suggests functions in transcription and chromatin regulation. *Genome Biol.* **14**, R119 (2013).
24. Kellinger, M.W. *et al.* 5-formylcytosine and 5-carboxylcytosine reduce the rate and substrate specificity of RNA polymerase II transcription. *Nat. Struct. Mol. Biol.* **19**, 831–833 (2012).
25. Wang, L. *et al.* Molecular basis for 5-carboxycytosine recognition by RNA polymerase II elongation complex. *Nature* **523**, 621–625 (2015).
26. Raiber, E.A. *et al.* 5-Formylcytosine alters the structure of the DNA double helix. *Nat. Struct. Mol. Biol.* **22**, 44–49 (2015).
27. Szulik, M.W. *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* **54**, 1294–1305 (2015).
28. Song, C.X. & He, C. Potential functional roles of DNA demethylation intermediates. *Trends Biochem. Sci.* **38**, 480–484 (2013).
29. Ko, M. *et al.* Impaired hydroxylation of 5-methylcytosine in myeloid cancers with mutant TET2. *Nature* **468**, 839–843 (2010).
30. Bachman, M. *et al.* 5-Hydroxymethylcytosine is a predominantly stable DNA modification. *Nat. Chem.* **6**, 1049–1055 (2014).
31. Shen, L. *et al.* Genome-wide analysis reveals TET- and TDG-dependent 5-methylcytosine oxidation dynamics. *Cell* **153**, 692–706 (2013).
32. Song, C.-X. *et al.* Genome-wide profiling of 5-formylcytosine reveals its roles in epigenetic priming. *Cell* **153**, 678–691 (2013).
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.* **21**, 1670–1676 (2011).
34. Bachman, M. *et al.* 5-Formylcytosine can be a stable DNA modification in mammals. *Nat. Chem. Biol.* **11**, 555–557 (2015).
35. Song, C.X., Yi, C. & He, C. Mapping recently identified nucleotide variants in the genome and transcriptome. *Nat. Biotechnol.* **30**, 1107–1116 (2012).
36. Plongthongkum, N., Diep, D.H. & Zhang, K. Advances in the profiling of DNA modifications: cytosine methylation and beyond. *Nat. Rev. Genet.* **15**, 647–661 (2014).
37. Booth, M.J., Raiber, E.A. & Balasubramanian, S. Chemical methods for decoding cytosine modifications in DNA. *Chem. Rev.* **115**, 2240–2254 (2015).
38. Booth, M.J. *et al.* Quantitative sequencing of 5-methylcytosine and 5-hydroxymethylcytosine at single-base resolution. *Science* **336**, 934–937 (2012).
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.* **25**, 679–684 (2011).
40. Pastor, W.A. *et al.* Genome-wide mapping of 5-hydroxymethylcytosine in embryonic stem cells. *Nature* **473**, 394–397 (2011).
41. Yu, M. *et al.* Base-resolution analysis of 5-hydroxymethylcytosine in the mammalian genome. *Cell* **149**, 1368–1380 (2012).
42. Stroud, H., Feng, S., Morey Kinney, S., Pradhan, S. & Jacobsen, S.E. 5-Hydroxymethylcytosine is associated with enhancers and gene bodies in human embryonic stem cells. *Genome Biol.* **12**, R54 (2011).
43. Szulwach, K.E. *et al.* Integrating 5-hydroxymethylcytosine into the epigenomic landscape of human embryonic stem cells. *PLoS Genet.* **7**, e1002154 (2011).
44. Szulwach, K.E. *et al.* 5-hmC-mediated epigenetic dynamics during postnatal neurodevelopment and aging. *Nat. Neurosci.* **14**, 1607–1616 (2011).
45. Lister, R. *et al.* Global epigenomic reconfiguration during mammalian brain development. *Science* **341**, 1237905 (2013).
46. Wang, L. *et al.* Programming and inheritance of parental DNA methylomes in mammals. *Cell* **157**, 979–991 (2014).
47. Shen, L. *et al.* Tet3 and DNA replication mediate demethylation of both the maternal and paternal genomes in mouse zygotes. *Cell Stem Cell* **15**, 459–470 (2014).
48. Guo, F. *et al.* Active and passive demethylation of male and female pronuclear DNA in the mammalian zygote. *Cell Stem Cell* **15**, 447–458 (2014).
49. Sun, Z. *et al.* A sensitive approach to map genome-wide 5-hydroxymethylcytosine and 5-formylcytosine at single-base resolution. *Mol. Cell* **57**, 750–761 (2015).
50. Song, C.X. *et al.* Selective chemical labeling reveals the genome-wide distribution of 5-hydroxymethylcytosine. *Nat. Biotechnol.* **29**, 68–72 (2011).
51. Booth, M.J., Marsico, G., Bachman, M., Beraldi, D. & Balasubramanian, S. Quantitative sequencing of 5-formylcytosine in DNA at single-base resolution. *Nat. Chem.* **6**, 435–440 (2014).
52. Lu, X. *et al.* Chemical modification-assisted bisulfite sequencing (CAB-Seq) for 5-carboxylcytosine detection in DNA. *J. Am. Chem. Soc.* **135**, 9315–9317 (2013).
53. Lu, X. *et al.* Base-resolution maps of 5-formylcytosine and 5-carboxylcytosine reveal genome-wide DNA demethylation dynamics. *Cell Res.* **25**, 386–389 (2015).
54. Wu, H., Wu, X., Shen, L. & Zhang, Y. Single-base resolution analysis of active DNA demethylation using methylase-assisted bisulfite sequencing. *Nat. Biotechnol.* **32**, 1231–1240 (2014).
55. Neri, F. *et al.* Single-base resolution analysis of 5-formyl and 5-carboxyl cytosine reveals promoter DNA methylation dynamics. *Cell Rep.* **10**, 674–683 (2015).
56. Ooi, S.K. *et al.* DNMT3L connects unmethylated lysine 4 of histone H3 to *de novo* methylation of DNA. *Nature* **448**, 714–717 (2007).
57. Boulard, M., Edwards, J.R. & Bestor, T.H. FBXL10 protects Polycomb-bound genes from hypermethylation. *Nat. Genet.* **47**, 479–485 (2015).