Current technological advances in mapping new DNA modifications

Rachel Amouroux[1], Kirsten R. McEwen[1] and Petra Hajkova[1†].

[1]Medical Research Council Clinical Sciences Centre, Imperial College London, London W12 0NN, United Kingdom.

[†]Correspondence: petra.hajkova@csc.mrc.ac.uk
Tel : 0044-2083838264

Abstract

The recent discovery of Tet (Ten eleven translocation) family of enzymes implicated in the chemical conversion of 5-methylcytosine (5mC) into 5-hydroxymethylcytosine (5hmC), 5-formylcytosine (5fC) and 5-carboxycytosine (5caC) has significantly enlarged the repertoire of known modifications present in the genomic DNA of higher eukaryotes. Considered as new epigenetic marks but also as DNA demethylation intermediates, intense research efforts have been directed towards deciphering of the biological function of Tet-driven DNA modifications. However, their low abundance in the mammalian genome, the relative similarity between their chemical structures and the predicted transient nature of 5fC and 5caC in genomic DNA make these modified DNA bases extremely challenging to study. The following review summarizes the techniques developed recently to quantify, profile and map 5hmC, 5fC and 5caC in a genomic context.

5-methylcytosine (5mC) has been considered a stable epigenetic mark ever since its identification several decades ago [1]. In contrast to the dynamic nature of histone modifications, DNA methylation patterns have shown to be stable and faithfully inherited through rounds of DNA replication in somatic cells. This attribute of DNA methylation has been mechanistically linked with phenotypic stability and maintenance of cell fate. Consequently, dynamic changes in DNA methylation patterns have been observed only during processes that involve cell fate reversal and global erasure of epigenetic information. *In vivo*, under physiological conditions, these processes are known to occur only twice during mammalian development: in the zygote immediately following fertilisation and in primordial germ cells (PGCs) [2]. Recent advances in our ability to manipulate cell fate and to reprogram cells *in vitro* [3] brought further mechanistic insights regarding the previously unappreciated dynamic character of DNA methylation patterns. Reprogramming of somatic cells or nuclei back to pluri- or totipotency through induced pluripotent stem cell (iPS) generation, cell fusion or somatic cell nuclear transfer (SCNT) is crucially connected with genome-wide changes in DNA methylation [2]. Additionally, the failure to efficiently reprogram DNA methylation patterns has been described as one of the barriers in the reprogramming process [4] The rapid and global changes in 5mC abundance and distribution observed during epigenetic reprogramming processes [5-8] and the anticipated key role of the underlying mechanism in the reprogramming efficiency have instigated an intensive search for potential molecular pathways implicated in DNA demethylation.

The pursuit of the DNA demethylation mechanism has recently re-gained momentum due to the discovery of new DNA bases in the mammalian genome [9, 10]. Resulting from the successive oxidation of 5mC by Tet1-3 enzymes, the newly identified 5-hydroxymethylcytosine (5hmC), 5-formylcytosine (5fC) and 5-carboxycytosine (5caC) bases have been subjected to intense research focus in attempt to determine their biological function (figure 1). The existence of higher oxidative products of 5mC has opened up a potential chemical route for the removal of 5mC and consequently various DNA demethylation pathways involving these oxidative intermediates have been suggested (figure 1). Amongst the proposed mechanisms, the only DNA demethylation pathway supported by clear biochemical evidence involves the recognition and cleavage of 5fC and 5caC by thymine-DNA glycosylase (TDG), followed by the activation of base excision repair (BER) to restore an unmodified DNA base [11]. Other proposed mechanisms involve deamination of 5hmC, decarboxylation of 5caC or direct removal of the hydroxymethyl group of 5hmC (figure 1) [11]. The oxidative products of 5mC are challenging to study due to their low abundance in the genome, the similarity between their chemical structures and the predicted transient nature of 5fC and 5caC. In consequence, the quantification and mapping

of these modified bases with sufficient accuracy and specificity has proven to be a technical feat. The following sections will review current methodologies recently developed to answer these questions.

Global detection and quantification of nucleoside variants: first step for the discovery of new bases

The presence of 5hmC in the genome of T-even bacteriophages was originally described in the 1950s [12], followed by reports of the detection of this modified base in animal DNA [13]. However, the recent discovery of the family of enzymes responsible for direct conversion of 5mC to 5hmC and the suggested role for 5hmC in DNA demethylation brought about a renewed interest in this cytosine modification.

Used for global assessment of DNA content and requiring a large amount of starting material, the thin-layer chromatography (TLC) technique has been greatly improved since its first use [14]. Based on the separation of compounds by capillary action on a stationary phase, this method can be combined with radioactive labelling to improve sensitivity. Pertaining to this, the two research teams implicated in the re-discovery of 5hmC used initially 1D and 2D-TLC [9, 10], confirming their finding subsequently by mass spectrometry. While initially used for separation of dC, 5mdC and 5hmdC, the optimization of buffer conditions has made the method recently suitable also for the detection of 5fC and 5caC [15-17].

Apart from the above mentioned chemical approaches, antibody-based techniques have been widely used for the detection and the semi-quantitative measurement of DNA modifications. Immunofluorescence staining using commercial or custom-made antibodies specifically recognising 5hmC, 5fC or 5caC has been frequently used, especially in instances where only a limited amount of material is available such as early mouse embryos (e.g. [18-21]). Although the specificity of the antibodies used is usually confirmed by dot blot and competition assays, the linearity, sensitivity and cross-reactivity under actual experimental conditions remain a serious concern. Moreover, the cross-comparison between different samples (for instance control versus treated) requires internal controls due to the inherent variability of signal obtained by immunofluorescent staining.

In order to increase the specificity and accuracy of 5hmC quantification, a number of new methods incorporated the use of T4-β-glucosyltransferase (T4-BGT). T4-BGT, an enzyme found in T-even phages, catalyses the covalent addition of uridine diphosphoglucose (UDP-

glucose) on the hydroxyl group of 5hmC to generate β-glucosyl-5-hydroxymethylcytosine (g5hmC) [22]. T4-BGT glucosylates 5hmC with high efficiency in a distributive, context-independent manner. The efficiency of this reaction is also not affected by 5hmC density, which is a common problem in antibody based 5hmC detection methods. The enzymatic conversion of 5hmC to g5hmC has been used in techniques for both the global quantification (using labelled UDP-Glucose [22, 23]) and mapping of 5hmC (TABseq, Aba-seq, discussed in the following section).

Ultimately, the gold standard used to identify and accurately quantify DNA modifications is provided by chromatographic separation of modified nucleosides (typically liquid chromatography (LC)), coupled with mass spectroscopy (MS). LC/MS delivers specificity by monitoring selected mass ion transitions and extreme sensitivity by enhanced signal-to-noise ratio. LC/MS has been successfully used to detect and quantify 5hmC, 5fC and 5caC in mouse embryonic stem cells (ESCs) and mammalian tissues [5, 9, 10, 16, 17, 24-26]. Despite its superior sensitivity and accuracy, one of the often underestimated drawbacks of the LC/MS approach is the occurrence of ion suppression. This phenomenon results from the presence of less volatile compounds such as salts, ion-pairing agents, endogenous molecules and metabolites, which perturb the ionization and thus affect the amount of charged ions that ultimately reach the detector of the mass spectrometer [27]. Either suppression or enhancement of the signal can result from co-elution of matrix components (i.e. enzyme digestion mix, salts, glycerol) with the compound of interest. The chromatographic conditions, including elution buffer composition and especially the ion-pairing agent present in the buffer can affect the signal intensity and signal-to-noise ratio. Additionally, the sample concentration can also influence the quantification of the amount of modified bases. As recently shown by Tang and colleagues, the signal response of 5hmC is significantly suppressed by an increasing amount of digested DNA loaded on the LC/MS system. This clearly demonstrates that poor sensitivity of a particular instrument or method cannot necessarily be counterbalanced by an increased amount of material loaded for analysis [28].

Several approaches can be implemented to address the effect of ion suppression. Isotope-labelled standards can be spiked into the samples to provide an internal control [24, 25]; the isotope-labelled “heavy” standard will be affected the same way as its “light” counterpart and thus can be used to monitor the loss/gain of signal caused by ion suppression. This technique has been used to accurately quantify 5hmC in different parts of the brain, confirming 5hmC as a new post-replicative mark [25]. It should be noted, however, that even the use of labelled internal standards cannot fully resolve the accuracy problems if ion

suppression is observed and hence one must carefully investigate the potential deleterious effects of various components such as the matrix on the final quantification.

Mapping of DNA modifications: towards the discovery of their functions

Determining whether or not 5hmC is a true epigenetic mark, an intermediate for DNA demethylation or a product of oxidative stress damage is a scientific challenge adopted by numerous teams. One of the key steps towards solving this question is to analyse the localisation and local abundance of these DNA modifications. Several new techniques designed to map 5hmC, 5fC and 5caC have recently emerged; most of these techniques have been commercialised and can be obtained as a kit at low cost. These are described below and summarized in figure 2.

5hmC quantification at specific genomic sites

GlucMS-PCR (glucosylation followed by methylation sensitive PCR) is used to interrogate 5mC and 5hmC levels at a particular CCGG genomic site. The method originally developed by Kinney *et al.* [29] combines treatment of genomic DNA with T4-BGT and restriction enzyme digest. The restriction isoschizomers used differ in their recognition and subsequent cleavage of modified CCGG sites: HpaII recognises only unmethylated sites, whereas MspI cleaves 5mC and 5hmC, but not 5ghmC. Following T4-BGT conversion and restriction digestion with HpaII and MspI carried out in parallel on both mock- and T4-BGT-treated samples, semi-quantitative analysis of end-point PCR products spanning the CCGG-containing restriction site allows an estimation of the relative amount of 5mC compared to 5hmC.

Low to medium resolution mapping approaches

In a similar manner to methylated DNA immunoprecipitation (meDIP), the hydroxy-meDIP (hmeDIP) technique coupled with whole-genome sequencing provides a relatively low resolution, but high-throughput approach to 5hmC mapping. The method has been optimised and used in mouse ESCs [30-33] to show that 5hmC is enriched at intermediate to high CpG-density regions, transcription-start sites and gene bodies. However, even though the 5hmC antibody shows high specificity compared to 5mC, meticulous studies have documented that the binding capacity of this antibody depends on the 5hmC density [34]; the antibody additionally displays off-target affinity for CA dinucleotides and simple repeats [35].

To avoid issues associated with antibody specificity towards 5hmC, methods using chemical modification or direct labelling of 5hmC have evolved. Rao's group generated an antibody recognizing cytosine 5-methylene sulfonate (CMS), the product of the reaction of sodium bisulfite with 5hmC [34]. Anti-CMS antibodies are far more specific than the anti-5hmC counterpart; they do not depend on 5hmC density and show much higher sensitivity in dot blot assays, allowing for more accurate immunoprecipitation experiments (CMS-IP) [34, 36]. Importantly, the presence of CMS does not affect PCR efficiency during library preparation except at very high, biologically irrelevant densities [34].

In an alternative method called GLIB-IP (glucosylation, periodate oxidation, biotinylation) also developed in Rao's lab, 5hmC is firstly glucosylated by T4-BGT and the glucose moiety is then oxidized with sodium periodate, which converts two hydroxyl groups to aldehydes. The aldehyde groups can then be chemically cross-linked with biotin using an aldehyde-reaction probe (ARP) followed by immunoprecipitation with streptavidin coupled beads. During PCR, the biotinylated adduct has only a minimal inhibitory effect and does not produce mutations [34, 37].

The use of chemically modified UDP-glucose in combination with T4-BGT has allowed researchers to be even more stringent and specific in their pull-down experiments. In the hmeSEAL (Selective Chemical Labeling and Capture) method [38], Song's group used an azide-modified glucose (UDP-6-N_3-glucose) and T4-BGT to selectively label 5hmC on DNA. The N_3 group allows chemical coupling with biotin in a reaction known as "click" chemistry (Huisgen cycloaddition). Quantification of modified 5hmC is possible through avidin-horseradish peroxidase (HRP) detection; alternatively, streptavidin pull-down and deep-sequencing analysis provides genome-wide distribution profiles. The latter has revealed that 5hmC is prominent in gene bodies of highly transcribed genes and seems to accumulate with age in mouse cerebellum [38].

In an alternative approach, Klungland and colleagues searched for enzymes recognizing g5hmC [39]. The genome of *Trypanosoma brucei* contains base J (β-glucosyl-5-hydroxymethyluracil). This glucosylated T derivative replaces about 1% of total thymine in the *T.brucei* genome and is specifically recognised by the J-binding protein JBP1. Interestingly, JBP1 also shows a strong affinity for g5hmC as these bases differ only by a single amine group. Klungland *et al*. describe a protocol in which JBP1 is covalently linked to magnetic beads and used for the pull-down of g5hmC-containing DNA. Although the resolution of this technique is relatively high (50bp), highly 5hmC-enriched sequences are

likely to be overrepresented due to increased pull-down efficiency [39]. Moreover, a comparative analysis between hmeDIP, hmeSEAL and JBP-1 affinity-based methods indicates that although results generated using hmeDIP and hmeSEAL are highly reproducible in detecting genome-wide patterns of 5hmC across different tissues, profiling of 5hmC distribution using JBP1 shows only poor overlap [35].

While considerable effort has been invested into the development of 5hmC mapping methods, there is a lack of techniques available for the detection and localization of 5fC and especially 5caC in a genomic context. This is mainly due to the very low abundance of these two bases (10 and 100 times lower than 5hmC, respectively). One of the methods described so far to profile 5fC relies on the chemical reaction (first described in [26]) of the aldehyde moiety present on 5fC with the oxyamine group of an ARP which contains biotin. This is followed by streptavidin based pull-down and genome-wide sequencing. Using this approach (called 5fC-DP-seq), 5fC enrichment in ESCs was found to follow a similar pattern as that observed for 5hmC (exons, promoters and CpG islands) [40].

More recently, He's group developed two different techniques for genome-wide profiling of 5fC. For profiling by pull-down (5fC-SEAL method), pre-existing 5hmC is first converted to g5hmC using T4-BGT. In a second step, 5fC is specifically reduced to 5hmC using sodium borohydrate ($NaBH_4$) and the newly generated 5hmC is tagged using UDP-6-N_3-glucose and selectively captured by "click" chemistry, as in the hmeSEAL method [41]. An alternative, single-base resolution mapping approach developed by this group is explained in the next section.

As described above, most of the techniques developed to regionally map 5hmC or 5hmC-oxidative products rely on chemical modification using either T4-BGT or ARP. In this context, it is important to note that in addition to 5hmC, T4-BGT also recognises and modifies 5hmU (5-hydroxymethyluracil), which is a product of either 5hmC deamination or T oxidation. Although 5hmU is mostly generated by oxidative stress and hence is present in only small quantities in genomic DNA, it can produce background signal in GLIB-IP and hmeSEAL experiments. Moreover, the symmetrical pattern of 5hmC seems to affect glucosylation efficiency, with hemi-hydroxymethylated CpG being less efficiently modified by T4-BGT. This substrate preference can thus potentially interfere with studies on 5hmC symmetry [22]. A final drawback of some of these methods is the known capacity of ARP to crosslink abasic sites as well as aldehyde moieties. Abasic sites naturally occur by hydrolysis of N-glycosidic bonds between the base and the phosphorus DNA backbone, leading to an average of 0.08-

0.25 abasic sites per 10^6bp per cell [42] potentially producing (albeit low) background signal. Further pros and cons of these methods are described in table 1.

Genome-wide mapping of DNA modifications at single base resolution

It is well documented that 5mC influences the chromatin landscape by recruiting readers (i.e. chromatin regulators) and thus exerts an effect on transcription. Similarly, 5hmC has been proposed to either mask 5mC in order to exclude these readers or to serve as a platform for the recruitment of 5hmC specific binders [43-46]. As affinity based strategies do not generally provide enough resolution to investigate 5hmC function/presence on specific sequences such as transcription factor binding sites or transcription start sites, there is a clear requirement for genome-wide mapping of 5hmC at single-base resolution.

Bisulfite-sequencing (BS-seq) relies on the chemically induced deamination of DNA yielding conversion of cytosine to uracil, whilst 5mC is refractory to such modification [47]. In consequence, as uracil is converted to T following PCR amplification, any remaining Cs represent original 5mC sites, allowing for direct 5mC mapping by sequencing. Although widely-used, this technique has several drawbacks. First, as 5hmC is also resistant to deamination during the bisulphite reaction, BS-seq cannot distinguish between these two DNA modifications. Additionally, 5fC and 5caC are indistinguishable from unmodified C, as they become deaminated and recognized as a T after bisulfite sequencing [48-50]. Second, the harsh conditions of the treatment induce DNA degradation can be problematic when dealing with a very low amount of starting material. Finally, primers and PCR conditions must be carefully optimised in order to avoid amplification bias towards C-rich or T-rich fragments. It should also be noted that this technique cannot be used to assess the global level of 5mC in the genome because of the presence of repetitive elements with unknown copy number. Additional techniques related to 5mC mapping and their drawbacks have been reviewed elsewhere [51]; we will hence further focus only on methods related to mapping of 5hmC, 5fC and 5caC, shown in figure 3.

Different strategies have been formulated to discriminate 5hmC and 5mC using modified BS-seq techniques. Tet-assisted bisulfite sequencing (TABseq) requires the initial protection of 5hmC by glucosylation using T4-BGT. The method subsequently relies on *in vitro* enzymatic oxidation of 5mC to 5caC using a recombinant Tet1 catalytic domain. Following bisulfite conversion, original 5mCs are converted to Ts and only 5hmC sites are read as Cs [52]. Compared to affinity-based strategies which conclude that 5hmC is enriched at CpG-dense

promoters, this method showed that 5hmC is mostly present on distal regulatory elements in human and mouse ESCs [50].

Oxidative BS-seq (oxBS-seq) discriminates between 5mC and 5hmC via highly selective chemical oxidation of 5hmC to 5fC using potassium perruthenate ($KRuO_4$). Subsequent bisulfite treatment causes 5fC to be deformylated and deaminated to U. Thus, any remaining Cs after oxBS-seq specifically represent 5mC [49, 53]. Direct comparison of results generated by oxBS-seq with traditional BS-seq can allow discrimination between 5mC and 5hmC.

Despite the extremely low abundance of 5fC and 5caC in genomic DNA, two new techniques have recently emerged to map these modified bases at single base resolution. In the CAB-seq method (chemical modification-assisted bisulfite sequencing), specific chemical modification using EDC (1-ethyl-3-[3-dimethylaminopropyl]-carbodiimide hydro-chloride) allows the protection of 5caC by forming an *o*-acylisourea reactive ester. After an EDC-coupling reaction, biotin can be added onto the modified 5caC for streptavidin-pull-down. This chemical conversion makes 5caC resistant to bisulfite-driven deamination, allowing 5caC mapping after bisulfite-sequencing by comparison between protected *versus* unprotected DNA. Importantly, the EDC-coupling reaction of 5caC does not introduce bias during PCR amplification [54]. This recently optimised technique has not been used so far to map 5caC in the mammalian genome.

For 5fC mapping, He & colleagues developed a similar technique (fCAB-seq) based on the specific protection of 5fC using hydroxylamine reduction, which prevents 5fC from undergoing bisulfite-mediated deamination. Consequently, the comparison of traditional BS-seq and fCAB-seq reveals the presence of 5fC at single base resolution [41].

All the above mentioned methods combine various chemical reactions with traditional bisulfite sequencing. The clear advantage of these approaches is the simultaneous detection of both unmodified and modified sequences, which provide a direct readout of the level of modification at a particular site (see table 1). However, as the levels of 5hmC, 5fC, 5caC modifications in genomic DNA are very low, reliable quantification by these methods requires very high coverage sequencing. This remains extremely costly for genomes of mammalian size. To circumvent this, some methods include an enrichment step which consequently allows higher sequence coverage for the genomic regions containing the modified base. This, however, prevents reliable quantification of the site specific modification level.

Considering the above mentioned constraints, alternative approaches to map oxidative products of Tet enzymes are of particular interest.

The unique specificity of the restriction endonuclease AbaSI to recognise 5ghmC has been harnessed by Zhen's group for AbaSI sequencing (Aba-seq) [55]. 5hmC present in genomic DNA is firstly converted to 5ghmC using T4-BGT. The modified DNA is then digested with AbaSI, generating DNA double-strand breaks (DSBs) 11-13bp 3' of 5ghmC sites. Biotinylated adaptors are next ligated to capture AbaSI-digested DNA and used for library preparation and sequencing. The presence of a C at the expected distance from the adaptor subsequently identifies the original 5hmC sites. Contrary to hmeDIP or TABseq, this method is not biased towards regions with high density or levels of 5hmC [55]. However, the method is extremely sensitive to the quality of the initial genomic DNA as the presence of DSBs caused by mechanical shearing (AbaSI independent) may lead to identification of false positive 5hmC sites.

The ultimate goal in mapping of DNA modifications is to "read" the presence of a modified nucleoside directly from the DNA molecule. In this context, the recently developed single-molecule real-time sequencing (SMRT-seq) method allows identification of epigenetic modifications on individual DNA molecules with single base resolution [56, 57]. For this third-generation sequencing, a single DNA polymerase molecule is monitored during incorporation of fluorescently labelled nucleotides into newly synthesised DNA. The incorporation of each new nucleotide generates a fluorescent pulse whose length and frequency yields information on the polymerase kinetics, reflecting a signature of the DNA structure and composition. In a first attempt to optimize this technique for the detection of DNA modifications, Flusberg *et al.* successfully detected N6-methyladenine, however differential kinetic signatures were missing for 5mC and 5hmC [56]. Coupled with chemical modification using T4-BGT enzyme and UDP-6-N_3-glucose, a modified version of SMRT-seq with a 5hmC enrichment procedure has been optimised, increasing the confidence of 5hmC assignments [57]. However, the field is still eagerly awaiting the validation of this method for 5hmC mapping in the mammalian genome. An obvious drawback of this technique is the cost of the instrument (PacBio *RS* High-Resolution Genetic Analyzer) and the relative low throughput (in comparison with standard next generation sequencing) precluding genome-wide studies.

Outlook and future technological challenges

The identification of new DNA modifications present in the mammalian genome and the emergence of novel technologies for their detection and quantification within a genomic context have raised new scientific questions and challenges.

The currently available methods each have shortcomings in sensitivity, large input material requirements or sequencing costs, limiting their use (table 1). As opposed to studies using *in vitro* systems of genome-wide DNA methylation remodelling, such as pluripotent stem cells [5], the study of *in vivo* demethylation in mouse zygotes and PGCs requires ultra-sensitive methods due to the limited availability of the starting material. Considerable efforts are thus required to improve and develop techniques suitable for small cell numbers. Furthermore, diverse chemical methods are currently required to probe the existence of individual DNA modifications. In this context, the commonly used BS-seq method does not distinguish between 5mC and 5hmC modifications; in view of this, caution should be taken when interpreting previously published data. Additional technological challenges will thus involve development of a single method capable of delivering information regarding all known DNA modifications simultaneously. In this context, the technological progress in the field of single molecule sequencing is eagerly awaited by the field.

Aside from the technological challenges, an important question concerns the underlying biological function of 5hmC, 5fC and 5caC, which still remains the subject of intense scientific debate. These modifications are considered as DNA demethylation intermediates, while also being recognized as the 6^{th}, 7^{th} and 8^{th} bases of the mammalian genome with potential biological functions distinct from the DNA demethylation pathway [11]. A key attribute of a demethylation intermediate would be a transient, highly dynamic behaviour, as is the case for oxidative DNA lesions. However a true epigenetic mark should be, like 5mC, heritable, stable and specifically recognized by molecular readers. Results from various biological systems including mouse zygotes and PGCs provide grounds to support either of these hypotheses [21, 58-61]. For example, numerous papers have suggested a role for 5hmC and further 5hmC oxidative products as intermediates during zygotic DNA demethylation [21, 60, 61]. On the other hand, mapping studies on these modifications in mouse and human ESCs clearly show their stable enrichment on specific genomic loci such as gene promoters, suggesting a distinct epigenetic function potentially linked with

transcriptional regulation. Using a proteomics approach, a set of proteins specifically recognizing 5hmC has been identified [43-46] with overrepresentation of repair-associated factors. While 5hmC is clearly a product of enzymatic TET-mediated oxidation of 5mC, it can also be considered as an oxidative lesion. Although rare, 5hmC and 5fC (but not 5caC) can be generated in the absence of TET enzymes by reactions involving either hydroxyl radical or one-electron oxidants (reviewed in [62]). In consequence, the oxidative stress generated during DNA modification profiling procedures should be accurately monitored and restrained using oxidant scavengers. These findings thus raise further questions on the underlying biological functions of 5hmC and 5fC.

Developing existing technologies and discovering novel approaches to assess the oxidative products of 5mC has driven intense research since their initial discovery. Continued efforts in this area will allow us to define the biological role of these intriguing modifications and their exact contribution to DNA demethylation processes. The recent discovery of Tet enzymes and their oxidative products may thus be a first step towards the discovery of other novel DNA base modifications participating in the complex epigenetic regulation of mammalian cells.

Acknowledgement

We would like to thank members of the Hajkova laboratory for helpful discussions. Work in the Hajkova laboratory is supported by the Medical Research Council (MC_US_A652_5PY70) and by funding from the EpigeneSys network (EU FP7) to P.H. P.H is a RISE 1 member of the EpigeneSys (EU FP7) network and a member of the EMBO Young Investigator Programme.

Conflict of interest

The authors have no conflict of interest to declare.

References

[1] Salomon R, Kaye AM, Herzberg M. Mouse nuclear satellite DNA: 5-methylcytosine content, pyrimidine isoplith distribution and electron microscopic appearance. J Mol Biol 1969; 43 (3)**:** 581-592.

[2] Hajkova P. Epigenetic reprogramming--taking a lesson from the embryo. Curr Opin Cell Biol 2010; 22 (3)**:** 342-350.

[3] Takahashi K, Yamanaka S. Induction of pluripotent stem cells from mouse embryonic and adult fibroblast cultures by defined factors. Cell 2006; 126 (4)**:** 663-676.

[4] Mikkelsen TS, Hanna J, Zhang X, Ku M, Wernig M, Schorderet P*, et al.* Dissecting direct reprogramming through integrative genomic analysis. Nature 2008; 454 (7200)**:** 49-55.

[5] Leitch HG, McEwen KR, Turp A, Encheva V, Carroll T, Grabole N*, et al.* Naive pluripotency is associated with global DNA hypomethylation. Nat Struct Mol Biol 2013; 20 (3)**:** 311-316.

[6] Hajkova P, Erhardt S, Lane N, Haaf T, El-Maarri O, Reik W*, et al.* Epigenetic reprogramming in mouse primordial germ cells. Mech Dev 2002; 117 (1-2)**:** 15-23.

[7] Oswald J, Engemann S, Lane N, Mayer W, Olek A, Fundele R*, et al.* Active demethylation of the paternal genome in the mouse zygote. Curr Biol 2000; 10 (8)**:** 475-478.

[8] Mayer W, Niveleau A, Walter J, Fundele R, Haaf T. Demethylation of the zygotic paternal genome. Nature 2000; 403 (6769)**:** 501-502.

[9] Kriaucionis S, Heintz N. The nuclear DNA base 5-hydroxymethylcytosine is present in Purkinje neurons and the brain. Science 2009; 324 (5929)**:** 929-930.

[10] Tahiliani M, Koh KP, Shen Y, Pastor WA, Bandukwala H, Brudno Y*, et al.* Conversion of 5-methylcytosine to 5-hydroxymethylcytosine in mammalian DNA by MLL partner TET1. Science 2009; 324 (5929)**:** 930-935.

[11] 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 (6)**:** 341-356.

[12] Wyatt GR, Cohen SS. The bases of the nucleic acids of some bacterial and animal viruses: the occurrence of 5-hydroxymethylcytosine. Biochem J 1953; 55 (5)**:** 774-782.

[13] Penn NW, Suwalski R, O'Riley C, Bojanowski K, Yura R. The presence of 5-hydroxymethylcytosine in animal deoxyribonucleic acid. Biochem J 1972; 126 (4)**:** 781-790.

[14] Grippo P, Iaccarino M, Rossi M, Scarano E. Thin-Layer Chromatography of Nucleotides, Nucleosides and Nucleic Acid Bases. Biochim Biophys Acta 1965; 95**:** 1-7.

[15] Shen L, Zhang Y. Enzymatic analysis of Tet proteins: key enzymes in the metabolism of DNA methylation. Methods Enzymol 2012; 512**:** 93-105.

[16] Ito S, Shen L, Dai Q, Wu SC, Collins LB, Swenberg JA*, et al.* Tet proteins can convert 5-methylcytosine to 5-formylcytosine and 5-carboxylcytosine. Science 2011; 333 (6047)**:** 1300-1303.

[17] He YF, Li BZ, Li Z, Liu P, Wang Y, Tang Q*, et al.* Tet-mediated formation of 5-carboxylcytosine and its excision by TDG in mammalian DNA. Science 2011; 333 (6047)**:** 1303-1307.

[18] Hajkova P, Jeffries SJ, Lee C, Miller N, Jackson SP, Surani MA. Genome-wide reprogramming in the mouse germ line entails the base excision repair pathway. Science 2010; 329 (5987)**:** 78-82.

[19] 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 (12)**:** 1670-1676.

[20] Salvaing J, Aguirre-Lavin T, Boulesteix C, Lehmann G, Debey P, Beaujean N. 5-Methylcytosine and 5-hydroxymethylcytosine spatiotemporal profiles in the mouse zygote. PLoS One 2012; 7 (5)**:** e38156.

[21] Wossidlo M, Nakamura T, Lepikhov K, Marques CJ, Zakhartchenko V, Boiani M*, et al.* 5-Hydroxymethylcytosine in the mammalian zygote is linked with epigenetic reprogramming. Nat Commun 2011; 2**:** 241.
[22] Terragni J, Bitinaite J, Zheng Y, Pradhan S. Biochemical characterization of recombinant beta-glucosyltransferase and analysis of global 5-hydroxymethylcytosine in unique genomes. Biochemistry 2012; 51 (5)**:** 1009-1019.
[23] Szwagierczak A, Bultmann S, Schmidt CS, Spada F, Leonhardt H. Sensitive enzymatic quantification of 5-hydroxymethylcytosine in genomic DNA. Nucleic Acids Res 2010; 38 (19)**:** e181.
[24] Globisch D, Munzel M, Muller M, Michalakis S, Wagner M, Koch S*, et al.* Tissue distribution of 5-hydroxymethylcytosine and search for active demethylation intermediates. PLoS One 2010; 5 (12)**:** e15367.
[25] Munzel M, Globisch D, Bruckl T, Wagner M, Welzmiller V, Michalakis S*, et al.* Quantification of the sixth DNA base hydroxymethylcytosine in the brain. Angew Chem Int Ed Engl 2010; 49 (31)**:** 5375-5377.
[26] Pfaffeneder T, Hackner B, Truss M, Munzel M, Muller M, Deiml CA*, et al.* The discovery of 5-formylcytosine in embryonic stem cell DNA. Angew Chem Int Ed Engl 2011; 50 (31)**:** 7008-7012.
[27] Annesley TM. Ion suppression in mass spectrometry. Clin Chem 2003; 49 (7)**:** 1041-1044.
[28] Tang Y, Chu JM, Huang W, Xiong J, Xing XW, Zhou X*, et al.* Hydrophilic material for the selective enrichment of 5-hydroxymethylcytosine and its liquid chromatography-tandem mass spectrometry detection. Anal Chem 2013; 85 (12)**:** 6129-6135.
[29] Kinney SM, Chin HG, Vaisvila R, Bitinaite J, Zheng Y, Esteve PO*, et al.* Tissue-specific distribution and dynamic changes of 5-hydroxymethylcytosine in mammalian genomes. J Biol Chem 2011; 286 (28)**:** 24685-24693.
[30] Ficz G, Branco MR, Seisenberger S, Santos F, Krueger F, Hore TA*, et al.* Dynamic regulation of 5-hydroxymethylcytosine in mouse ES cells and during differentiation. Nature 2011; 473 (7347)**:** 398-402.
[31] Xu Y, Wu F, Tan L, Kong L, Xiong L, Deng J*, et al.* Genome-wide regulation of 5hmC, 5mC, and gene expression by Tet1 hydroxylase in mouse embryonic stem cells. Mol Cell 2011; 42 (4)**:** 451-464.
[32] Wu H, D'Alessio AC, Ito S, Wang Z, Cui K, Zhao K*, 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 (7)**:** 679-684.
[33] Williams K, Christensen J, Pedersen MT, Johansen JV, Cloos PA, Rappsilber J*, et al.* TET1 and hydroxymethylcytosine in transcription and DNA methylation fidelity. Nature 2011; 473 (7347)**:** 343-348.
[34] Pastor WA, Pape UJ, Huang Y, Henderson HR, Lister R, Ko M*, et al.* Genome-wide mapping of 5-hydroxymethylcytosine in embryonic stem cells. Nature 2011; 473 (7347)**:** 394-397.
[35] Thomson JP, Hunter JM, Nestor CE, Dunican DS, Terranova R, Moggs JG*, et al.* Comparative analysis of affinity-based 5-hydroxymethylation enrichment techniques. Nucleic Acids Res 2013; 41 (22)**:** e206.
[36] Huang Y, Pastor WA, Zepeda-Martinez JA, Rao A. The anti-CMS technique for genome-wide mapping of 5-hydroxymethylcytosine. Nat Protoc 2012; 7 (10)**:** 1897-1908.
[37] Pastor WA, Huang Y, Henderson HR, Agarwal S, Rao A. The GLIB technique for genome-wide mapping of 5-hydroxymethylcytosine. Nat Protoc 2012; 7 (10)**:** 1909-1917.
[38] Song CX, Szulwach KE, Fu Y, Dai Q, Yi C, Li X*, et al.* Selective chemical labeling reveals the genome-wide distribution of 5-hydroxymethylcytosine. Nat Biotechnol 2011; 29 (1)**:** 68-72.
[39] Robertson AB, Dahl JA, Ougland R, Klungland A. Pull-down of 5-hydroxymethylcytosine DNA using JBP1-coated magnetic beads. Nat Protoc 2012; 7 (2)**:** 340-350.

[40] Raiber EA, Beraldi D, Ficz G, Burgess HE, Branco MR, Murat P*, et al.* Genome-wide distribution of 5-formylcytosine in embryonic stem cells is associated with transcription and depends on thymine DNA glycosylase. Genome Biol 2012; 13 (8)**:** R69.
[41] Song CX, Szulwach KE, Dai Q, Fu Y, Mao SQ, Lin L*, et al.* Genome-wide profiling of 5-formylcytosine reveals its roles in epigenetic priming. Cell 2013; 153 (3)**:** 678-691.
[42] Cadet J, Delatour T, Douki T, Gasparutto D, Pouget JP, Ravanat JL*, et al.* Hydroxyl radicals and DNA base damage. Mutat Res 1999; 424 (1-2)**:** 9-21.
[43] Yildirim O, Li R, Hung JH, Chen PB, Dong X, Ee LS*, et al.* Mbd3/NURD complex regulates expression of 5-hydroxymethylcytosine marked genes in embryonic stem cells. Cell 2011; 147 (7)**:** 1498-1510.
[44] Spruijt CG, Gnerlich F, Smits AH, Pfaffeneder T, Jansen PW, Bauer C*, et al.* Dynamic readers for 5-(hydroxy)methylcytosine and its oxidized derivatives. Cell 2013; 152 (5)**:** 1146-1159.
[45] Iurlaro M, Ficz G, Oxley D, Raiber EA, Bachman M, Booth MJ*, et al.* A screen for hydroxymethylcytosine and formylcytosine binding proteins suggests functions in transcription and chromatin regulation. Genome Biol 2013; 14 (10)**:** R119.
[46] Mellen M, Ayata P, Dewell S, Kriaucionis S, Heintz N. MeCP2 binds to 5hmC enriched within active genes and accessible chromatin in the nervous system. Cell 2012; 151 (7)**:** 1417-1430.
[47] Olek A, Oswald J, Walter J. A modified and improved method for bisulphite based cytosine methylation analysis. Nucleic Acids Res 1996; 24 (24)**:** 5064-5066.
[48] Huang Y, Pastor WA, Shen Y, Tahiliani M, Liu DR, Rao A. The behaviour of 5-hydroxymethylcytosine in bisulfite sequencing. PLoS One 2010; 5 (1)**:** e8888.
[49] Booth MJ, Branco MR, Ficz G, Oxley D, Krueger F, Reik W*, et al.* Quantitative sequencing of 5-methylcytosine and 5-hydroxymethylcytosine at single-base resolution. Science 2012; 336 (6083)**:** 934-937.
[50] Yu M, Hon GC, Szulwach KE, Song CX, Zhang L, Kim A*, et al.* Base-resolution analysis of 5-hydroxymethylcytosine in the mammalian genome. Cell 2012; 149 (6)**:** 1368-1380.
[51] Shanmuganathan R, Basheer NB, Amirthalingam L, Muthukumar H, Kaliaperumal R, Shanmugam K. Conventional and nanotechniques for DNA methylation profiling. J Mol Diagn 2013; 15 (1)**:** 17-26.
[52] Yu M, Hon GC, Szulwach KE, Song CX, Jin P, Ren B*, et al.* Tet-assisted bisulfite sequencing of 5-hydroxymethylcytosine. Nat Protoc 2012; 7 (12)**:** 2159-2170.
[53] Booth MJ, Ost TW, Beraldi D, Bell NM, Branco MR, Reik W*, et al.* Oxidative bisulfite sequencing of 5-methylcytosine and 5-hydroxymethylcytosine. Nat Protoc 2013; 8 (10)**:** 1841-1851.
[54] Lu X, Song CX, Szulwach K, Wang Z, Weidenbacher P, Jin P*, et al.* Chemical modification-assisted bisulfite sequencing (CAB-Seq) for 5-carboxylcytosine detection in DNA. J Am Chem Soc 2013; 135 (25)**:** 9315-9317.
[55] Sun Z, Terragni J, Borgaro JG, Liu Y, Yu L, Guan S*, et al.* High-resolution enzymatic mapping of genomic 5-hydroxymethylcytosine in mouse embryonic stem cells. Cell Rep 2013; 3 (2)**:** 567-576.
[56] Flusberg BA, Webster DR, Lee JH, Travers KJ, Olivares EC, Clark TA*, et al.* Direct detection of DNA methylation during single-molecule, real-time sequencing. Nat Methods 2010; 7 (6)**:** 461-465.
[57] Song CX, Clark TA, Lu XY, Kislyuk A, Dai Q, Turner SW*, et al.* Sensitive and specific single-molecule sequencing of 5-hydroxymethylcytosine. Nat Methods 2012; 9 (1)**:** 75-77.
[58] Hackett JA, Sengupta R, Zylicz JJ, Murakami K, Lee C, Down TA*, et al.* Germline DNA demethylation dynamics and imprint erasure through 5-hydroxymethylcytosine. Science 2013; 339 (6118)**:** 448-452.
[59] Yamaguchi S, Shen L, Liu Y, Sendler D, Zhang Y. Role of Tet1 in erasure of genomic imprinting. Nature 2013; 504 (7480)**:** 460-464.

[60] Gu TP, Guo F, Yang H, Wu HP, Xu GF, Liu W*, et al.* The role of Tet3 DNA dioxygenase in epigenetic reprogramming by oocytes. Nature 2011; 477 (7366)**:** 606-610.
[61] Iqbal K, Jin SG, Pfeifer GP, Szabo PE. Reprogramming of the paternal genome upon fertilization involves genome-wide oxidation of 5-methylcytosine. Proc Natl Acad Sci U S A 2011; 108 (9)**:** 3642-3647.
[62] Cadet J, Wagner JR. TET enzymatic oxidation of 5-methylcytosine, 5-hydroxymethylcytosine and 5-formylcytosine. Mutat Res 2013.

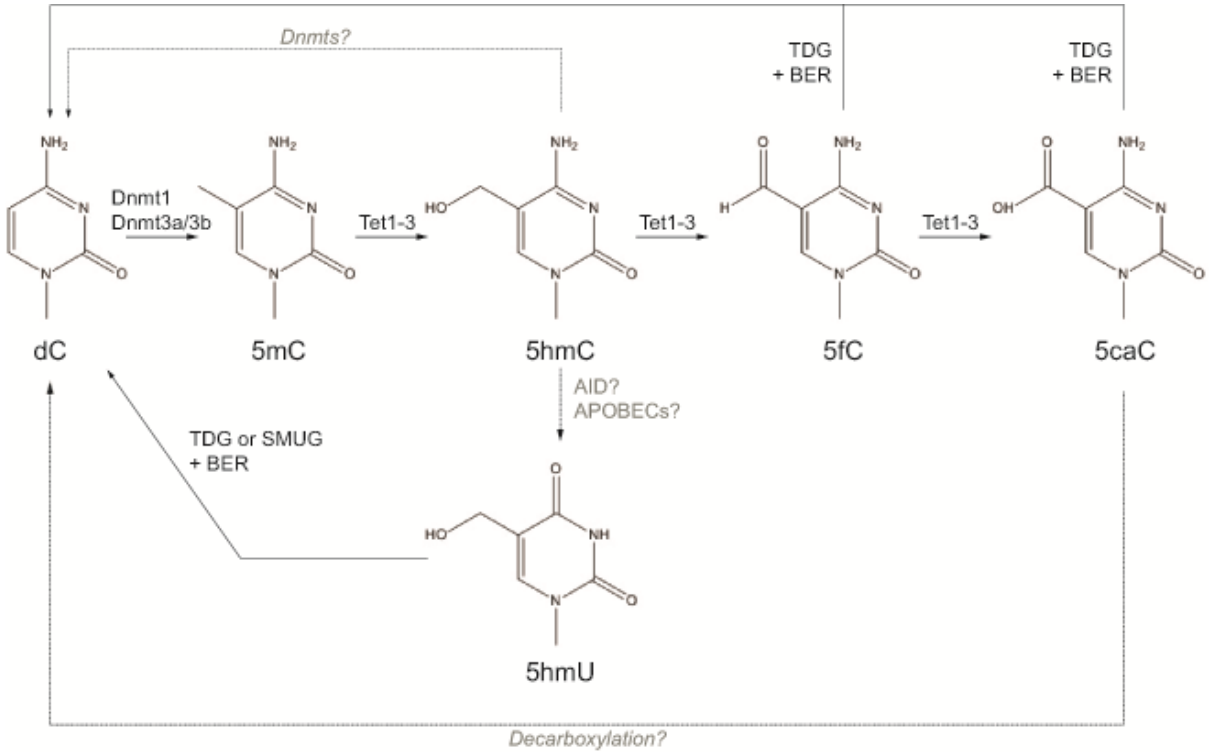

Figure 1. Conversion of 5-methylcytosine (5mC) to 5-hydroxymethylcytosine (5hmC), 5-formylcytosine (5fC) and 5-carboxycytosine (5caC) by Tet enzymes (Tet1-3) with known and putative mechanisms for DNA demethylation. TDG, thymine-DNA glycosylase; BER, base excision repair; SMUG, single-strand selective monofunctional uracil DNA glycosylase; AID, activation-induced cytidine deaminase; APOBEC, apolipoprotein B mRNA editing enzyme, catalytic polypeptide; Dnmt, DNA (cytosine-5-)-methyltransferase; Tet, Ten eleven translocation; dC, deoxycytosine; 5hmU, 5-hydroxymethyluracil.

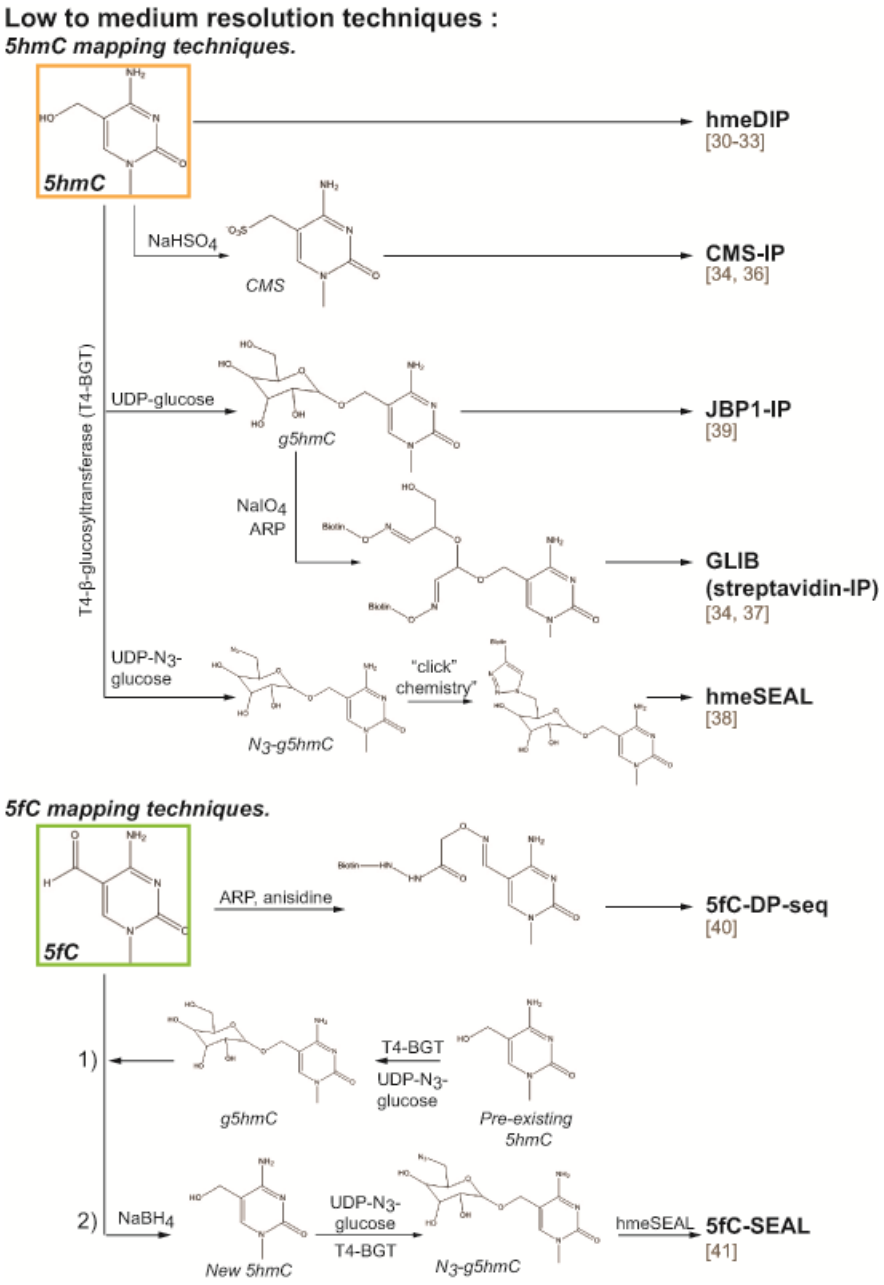

Figure 2. Schematic of the recent techniques used for the profiling of 5-hydroxymethylcytosine (5hmC) and 5-formylcytosine (5fC) at low to medium resolution. hmeDIP, hydroxymethylated DNA immunoprecipitation; CMS-IP, cytosine 5-methylene sulfonate immunoprecipitation; JBP1-IP, J-binding protein immunoprecipitation; GLIB-IP, glucosylation, periodate oxidation, biotinylation immunoprecipitation; hmeSEAL, hydroxymethyl selective chemical labeling and capture; 5fC-DP-seq, 5-formylcytosine DNA pulldown sequencing; 5fC-SEAL, 5fC-selective chemical labeling and capture; g5hmC, β-glucosyl-5-hydroxymethylcytosine; ARP, aldehyde-reaction probe; T4-BGT, T4-β-glucosyltransferase.

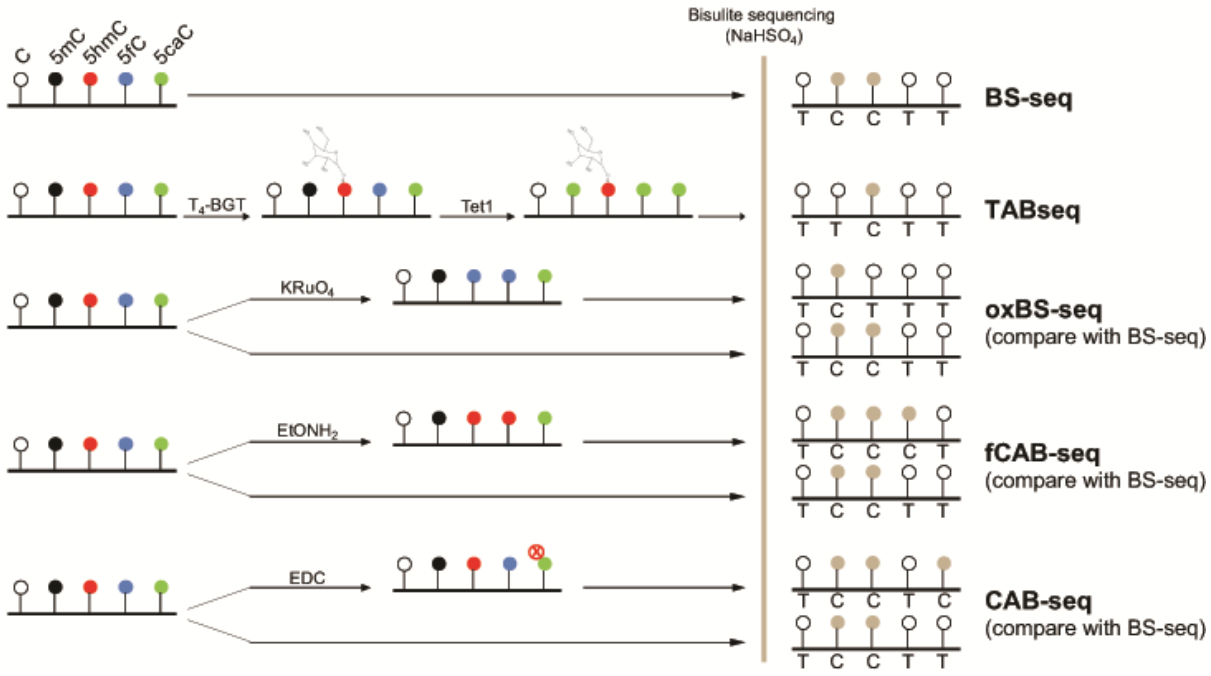

Figure 3. Genome-wide methods developed to map DNA modifications at single-base resolution. 5mC, 5-methylcytosine; 5hmC, 5-hydroxymethylcytosine; 5fC, 5-formylcytosine; 5caC, 5-carboxycytosine; T4-BGT, T4-β-glucosyltransferase; BS-seq, bisulfite sequencing; TABseq, Tet-assisted bisulfite sequencing; oxBS-seq, oxidative bisulfite sequencing; fCAB-seq, 5-formylcytosine chemical modification-assisted bisulfite sequencing; CAB-seq, chemical modification-assisted bisulfite sequencing.

	DNA modifications	Outcome	Time	Cost	Amount of material	Pros	Cons	Ref
1D, 2D-TLC	5mC, 5hmC, 5fC, 5caC, unknown	*global*	◷◷	$	ug DNA	Allows discovery of new modified bases.	Separation of new modified bases is dependent on buffer conditions. Requires subsequent confirmation by LC/MS.	[9, 10, 14-17]
LC-MS	5mC, 5hmC, 5fC, 5caC, unknown	*global*	◷	$$$$	ng DNA	Highly sensitive and specific, provides quantitative information (broad linear range). Discovery of new DNA modifications possible.	Time consuming method optimisation, ion suppression. High instrument cost.	[5, 9, 10, 16, 17, 24-26]
IF	5mC, 5hmC, 5fC, 5caC	*global*	◷	$	single cell	Relative quantification on single cells.	Dependent on antibody specificity, sensitivity, saturation and linear range. Semi-quantitative.	[18-21]
GlucMS-PCR	5hmC	*single-locus*	◷	$	ug DNA	Quick assessment of presence and level of DNA modifications.	Very low throughput: only 1 CCGG site examined at a time.	[29]
hmeDIP	5hmC	*genome-wide, low-medium resolution*	◷◷	$$	ug DNA	Simple protocol, provides genome-wide information.	Critically depends on antibody specificity and sensitivity. Efficiency of antibody precipitation increases with epitope concentration. Unspecific pull-down of CA and CT-rich sequences.	[30-33]
GLIB-IP	5hmC	*genome-wide, low-medium resolution*	◷◷	$$	ug DNA	Good linearity of response to epitope concentration. Precipitation of fragments containing a single 5hmC.	Several enzymatic and chemical modification steps. Higher background precipitation than CMS-IP. Underrepresentation of 5hmC-dense sequences due to slight inhibitory effect on PCR reaction. Potential background due to 5hmU.	[34, 37]
hmeSEAL	5hmC	*genome-wide, low-medium resolution*	◷◷	$$	ug DNA	High specificity of the CLICK reaction. Can label 5hmC with any tag (HRP, biotin, fluorescent).	Need chemical modifications. CLICK chemistry might affect downstream PCR reaction. Potential background due to 5hmU.	[38]
CMS-IP	5hmC	*genome-wide, low-medium resolution*	◷◷	$$	ug DNA	Higher specificity and sensitivity, less dependent on epitope concentration than 5hmC antibody.	Chemical modifications required. Not efficient for pulling-down sequences containing single 5hmC.	[34, 36]
JBP1-IP	5hmC	*genome-wide, medium resolution*	◷◷	$$	ug DNA	Good resolution (50bp).	Relies on protein specificity, affinity increases with epitope concentration. Results show poor overlap with hmeDIP and hmeSEAL. Potential background due to 5hmU.	[39]
5fC-SEAL	5fC	*genome-wide, low-medium resolution*	◷◷	$$	ug DNA	Highly specific and sensitive without 5fC density bias.	Chemical modifications required. CLICK chemistry might affect downstream PCR reaction. Potential background due to 5hmU.	[41]
5fC-DP-seq	5fC	*genome-wide, low-medium resolution*	◷◷	$$	ug DNA	1st method developed for 5fC pulldown. Chemistry specific for 5fC.	ARP cross-reacts with abasic sites.	[40]
BS-seq	5mC+5hmC	*genome-wide, single-bp resolution*	◷◷	$$	single cell	Whole genome coverage.	Does not distinguish between 5mC and 5hmC, dC and 5fC/5CaC.	[47]
SMRT-seq	5mC, 5hmC, mA	*genome-wide, single-bp resolution*	◷	$$$$	ug DNA	Only method available for the identification of different DNA modifications on a single DNA molecule.	Requires chemical modification (BGT, N_3-UDP-glucose) and enrichment of 5hmC-containing DNA. High instrument cost. Relatively low throughput.	[56, 57]
TABseq	5hmC	*genome-wide, single-bp resolution*	◷◷◷	$$	ug DNA	No enrichment procedure.	Relies on BGT and Tet1 activity (which is difficult to purify): high rate of false positives (5%). Requires very high read coverage.	[50, 52]
OxBS-seq	5hmC	*genome-wide, single-bp resolution*	◷◷	$$	ng-ug DNA	Whole genome coverage.	Chemical modification required (induction of oxidative DNA damage?). Requires substraction with normal BS-seq (5mC+5hmC) therefore very high read coverage needed. Lots of purification and buffer exchange steps.	[49, 53]
Aba-seq	5hmC	*genome-wide, single-bp resolution*	◷◷	$$	ng DNA	Simple library construction, good accuracy at low 5hmC density sites.	Enzymatic properties of AbaSI (different cutting efficiency depending on target sites). Complicated data analysis.	[55]
CAB-seq	5caC	*genome-wide, single-bp resolution*	◷◷	$$	ug DNA	Only method so far to identify 5caC at a single bp level.	Enrichment step required, biotin needs to be removed prior to high-throughput sequencing.	[54]
fCAB-seq	5fC	*genome-wide, single-bp resolution*	◷◷	$$	ug DNA	No enrichment procedure.	Requires chemical step and high coverage for comparison with BS-seq.	[41]

Table 1. Summary of recent methods optimised for the detection, quantification and mapping at low, medium or single-base resolution of Tet-dependent DNA modifications. TLC, thin-layer chromatography; LC-MS, liquid chromatography-mass spectroscopy; IF, immunofluorescence; GlucMS-PCR, glucosylation methylation sensitive PCR; hmeDIP, hydroxymethylated DNA immunoprecipitation; GLIB-IP, glucosylation, periodate oxidation, biotinylation immunoprecipitation; hmeSEAL, hydroxymethyl selective chemical labeling and capture; CMS-IP, cytosine 5-methylene sulfonate immunoprecipitation; JBP1-IP, J-binding protein immunoprecipitation; 5fC-SEAL, 5fC-selective chemical labeling and capture; 5fC-DP-seq, 5fC-DNA pulldown sequencing; BS-seq, bisulfite sequencing; SMRT-seq, single-molecule real-time sequencing; TAB-seq, Tet-assisted bisulfite sequencing; OxBS-seq, oxidative bisulfite sequencing; Aba-seq, AbaSI sequencing; CAB-seq, chemical modification-assisted bisulfite sequencing; fCAB-seq, 5-formylcytosine chemical modification-assisted bisulfite sequencing; 5mC, 5-methylcytosine; 5hmC, 5-hydroxymethylcytosine; 5fC, 5-formylcytosine; 5caC, 5-carboxycytosine; mA, methyladenine; BGT, β-glucosyltransferase.