Published in final edited form as:

Nat Chem. 2014 May ; 6(5): 435–440. doi:10.1038/nchem.1893.

Quantitative sequencing of 5-formylcytosine in DNA at single-base resolution

Michael J. Booth¹, Giovanni Marsico², Martin Bachman², Dario Beraldi², and Shankar Balasubramanian^{1,2}

¹Department of Chemistry, University of Cambridge, Cambridge, CB2 1EW, UK

²Cancer Research UK, Cambridge Institute, Li Ka Shing Centre, Robinson Way, Cambridge, CB2 0RE, UK

Abstract

The cytosine modifications 5-hydroxymethylcytosine (5hmC) and 5-formylcytosine (5fC) were recently found to exist in the genomic DNA of a wide range of mammalian cell types. It is important to now understand their role in normal biological function and disease. Here we introduce reduced bisulfite sequencing (redBS-Seq), a method to quantitatively decode 5fC in DNA at singlebase resolution, based on a selective chemical reduction of 5fC to 5hmC followed by bisulfite treatment. Following extensive validation on synthetic and genomic DNA, we combined redBS-Seq and oxidative bisulfite sequencing (oxBS-Seq) to generate the first combined genomic map of 5-methylcytosine, 5hmC and 5fC in mouse embryonic stem cells. Our experiments revealed that 5fC is present at relatively high levels in certain genomic locations, in comparison to 5hmC and 5mC. The combination of these chemical methods can quantify and precisely map these three cytosine derivatives in the genome and will help provide insights into their function.

Introduction

The epigenetic feature 5-methylcytosine (5mC) that occurs at CpG (cytosinephosphate-guanine) dinucleotides¹, was the only known enzymatically synthesised DNA base modification in the mammalian genome prior to 2009. Recently, three additional cytosine modifications were found to exist in mammalian DNA; 5-hydroxymethylcytosine (5hmC)^{2, 3}, 5-formylcytosine (5fC)^{4, 5} and 5-carboxylcytosine (5caC)⁵. 5hmC and 5fC have been found in many cell types, including all the major organs (for instance brain, heart, lung and kidney) and embryonic stem (ES) cells^{4, 5}, however 5caC has only been detected in a few cell types at 10 times lower levels than 5fC^{5, 6}. The ten-eleven translocation (TET) enzymes catalyse the sequential oxidation of 5mC to 5hmC, 5fC and 5caC respectively³. This oxidation pathway has been proposed as a potential route for the demethylation of 5mC

Author Contributions: M.J.B. and S.B. co-invented the redBS-Seq method. M.J.B. and S.B. conceived the experiments with contributions from all authors. M.J.B. and M.B. performed experimental work. G.M. and D.B. performed bioinformatics analysis. M.J.B. and S.B. wrote the manuscript with contributions from all authors.

Competing financial interests: M.J.B., and S.B. are shareholders in and M.J.B. and D.B. are consultants for Cambridge Epigenetix, Ltd. S.B. is an advisor to and shareholder of Illumina, Inc.

back to cytosine (C), for example by the action of thymine DNA glycosylase (TDG), which can excise 5fC and 5caC from DNA as part of the base excision repair (BER) pathway to restore the C-nucleotide⁷⁻⁹. There is also emerging data to suggest that the various chemical modifications may be selectively recognised in the DNA major groove by proteins. Binding of methyl-CpG-binding protein 2 (MeCP2) to 5hmC has been proposed as an epigenetic transcriptional regulator in neurons¹⁰. Furthermore, mutations in the DNA binding domain of MeCP2 seen in Rett syndrome, a neurodevelopmental disorder, preferentially impact 5hmC binding over 5mC binding¹⁰. The conversion of 5hmC to 5fC has been proposed as a distinct signal for DNA demethylation, due to the enriched binding of DNA-repair-associated proteins to 5fC¹¹.

To fully understand the importance of these modifications in detail, it is necessary to accurately detect them in genomic DNA, ideally at single base resolution. Affinity based methods that exploit either a 5fC-specific antibody¹², chemical reactivity with a hydroxylamine probe¹³ or the use of β -glucosyltransferase¹⁴ have been developed to enrich and sequence 5fC-containing DNA fragments to map the position of 5fC in genomic DNA. However, such approaches are low resolution (typically 100s to 1000s of bases) and non-linear in their quantitation.

Quantitative single base resolution sequencing of 5mC has been performed using sodium bisulfite. Treatment of DNA with sodium bisulfite leads to rapid deamination of the C bases to uracil (U)^{15, 16}. These converted Cs are subsequently read as thymines (T) when the DNA is sequenced, owing to the change in the Watson-Crick hydrogen bonding pattern essential for base pairing. 5mC is resistant to bisulfite conversion and is hence read as a C, which enables quantitative discrimination between unmodified Cs and 5mC at single base resolution¹⁷. BS-Seq has been widely used to map patterns of DNA methylation in a variety of organisms.

Treatment of 5hmC with bisulfite results in a stable cytosine-5-methylsulfonate (CMS) adduct that, like 5mC, is resistant to deamination and is therefore read as a cytosine during DNA sequencing¹⁸. 5mC and 5hmC cannot therefore be distinguished by bisulfite conversion alone. We recently developed oxidative bisulfite sequencing (oxBS-Seq) to quantitatively detect 5mC and 5hmC at single-base resolution^{19, 20}. OxBS-Seq discriminates between 5mC and 5hmC via a selective chemical oxidation of 5hmC to 5fC. 5hmC reads as a C following bisulfite treatment, however 5fC deformylates and deaminates to form U under bisulfite conditions¹⁹. OxBS-Seq exploits this different in reactivity between 5hmC and 5fC towards bisulfite. The only base that is not deaminated and therefore read as a C after oxBS-Seq is 5mC. Assessment of 5hmC levels and positions in the genome are then achieved computationally by comparing a BS-Seq experiment to an oxBS-Seq experiment^{19, 20}. Independently, an enzymatic method has also been developed by He and co-workers to sequence 5hmC, TET-Assisted Bisulfite Sequencing (TAB-Seq). In TAB-Seq, a combination of 5hmC glucosylation, 5mC oxidation to 5caC and bisulfite treatment is used to resolve 5hmC and 5mC at single base resolution²¹.

Here, we introduce a new method, reduced bisulfite sequencing (redBS-Seq), as a practical chemical method to quantitatively detect 5fC at single-base resolution. After validating

redBS-Seq on synthetic and genomic DNA, we have exploited its potential to produce the first quantitative, single-base resolution map of 5mC, 5hmC and 5fC from mouse embryonic stem (mES) cells through the combined use of BS-Seq, oxBS-Seq and redBS-Seq. The data provide insights into the relationship between 5mC, 5hmC and 5fC at sites in the genome and demonstrated that although the global levels of 5fC are very low, it is present at levels comparable to 5mC and 5hmC at specific sites.

Results and Discussion

5fC reduction in single stranded DNA

5fC deformylates then deaminates to U upon treatment with bisulfite to be read as T¹⁹, while 5hmC converts to (CMS) to be read as C¹⁸. Using a specific chemical reduction of 5fC to 5hmC in DNA, prior to bisulfite treatment, would lead to all 5fC bases being read as C. By comparing the output of reduced bisulfite sequencing (redBS-Seq, where 5fC reads as C) with that of BS-Seq (where 5fC reads as T) would enable the elucidation of 5fC quantitatively and at single-base resolution (Fig. 1). We chose to use the reductant sodium borohydride, as it is a small water-soluble molecule and DNA is stable in its presence²². The reactivity of sodium borohydride was carefully evaluated on 9mer single stranded DNA (ssDNA) oligonucleotide initially, as the bases would be more accessible to the reductant to test for non-specific reactions. A specific chemical reduction of 5fC to 5hmC using aqueous sodium borohydride was optimised on this ssDNA (Fig. 2a). This reduction does not result in any detectable reaction of any other bases in ssDNA, as determined by HPLC (Supplementary Fig. S1). Furthermore, sodium borohydride has been shown to reduce 5fC to 5hmC in DNA^{5, 14}.

Reduction and bisulfite treatment of 5fC in double stranded DNA

Reaction with sodium borohydride was found to reduce 5fC to 5hmC in the context of a 100mer double stranded DNA (dsDNA) fragment with an 80% recovered yield of 5hmC and no remaining 5fC, as measured by mass spectrometry of the 2'-deoxynucleosides (Fig. 2b). This suggested that the reagent could approach the formyl group in the context of the DNA major groove largely unhindered. Furthermore, reduction directly on dsDNA avoids the need for a denaturation step to liberate ssDNA prior to reduction. Upon reduction of 5fC containing DNA there was no resulting DNA fragmentation or loss of total dsDNA, as assessed by agarose gel electrophoresis and absorbance after dsDNA flurophore binding (Supplementary Fig. S2), which is important for ultimately sequencing small quantities of DNA. 5fC reads as a T following BS-Seq by Sanger sequencing (Fig. 2c), however by performing the chemical reduction followed by a sodium bisulfite treatment (redBS-Seq) we specifically converted 5fC to 5hmC then CMS to convert the reading of 5fC to C by Sanger sequencing (Fig. 2d). Furthermore, all other cytosine bases (C, 5mC, 5hmC and 5caC) have the same read-out with BS-Seq and redBS-Seq, as measured using Sanger sequencing (Supplementary Fig. S3 and S4). This demonstrates that the redBS-Seq method specifically modifies the Watson-Crick base-pairing pattern of only 5fC.

We next quantified the conversion efficiencies of 5fC and 5hmC following reduction then bisulfite treatment (redBS-Seq) in synthetic control dsDNA containing 5fC using Illumina

sequencing, which digitally counts the sequence of each molecule in a pool (Fig. 2d). The percentage of 5hmC and 5fC that read as either C or T following BS-Seq or redBS-Seq was therefore accurately quantified. We used dsDNA that contained 5fC and 5hmC in multiple contexts (two CpGs or one CpG and one non-CpG) to obtain an accurate value of conversion for each base and rule out any sequence bias in conversion. Following BS-Seq, the majority of 5fC deaminates to U and only 14% of residual 5fC reads as a C, whereas following redBS-Seq 97% of 5fC reads as a C. The level of 5fC at a single base would be calculated as the difference between these two methods ($97\% - 14\% = 83\%$). Therefore, the measured level of 5fC at each site would be underestimating the actual level by a factor of 17% (Supplementary Table 1). It is this large difference in conversion efficiencies for 5fC between the two sequencing methods that enables quantitative single-base resolution sequencing. There was little change in 5hmC conversion between the two treatments. The small observed conversion of 5hmC to U (12-16%) is due to the known contamination of the commercially supplied d5hmCTP with dCTP²³, which artificially increases the observed conversion of 5hmC to U.

Reduced bisulfite sequencing of genomic DNA

To validate the utility of the redBS-Seq method on genomic DNA, we chose to study mouse embryonic stem (mES) cells, as they are known to contain measurable levels of 5hmC and 5fC^{4, 5}. Initially, mass spectrometry was used to calculate the global percentage levels of 5hmC and 5fC over the total level of all Cs, which were found to be $0.0014 \pm 0.0003\%$ 5fC and $0.055 \pm 0.008\%$ 5hmC (mean $\pm$ standard deviation, $n=3$). Secondly, we subjected this mES cell DNA to a sodium borohydride reduction, and by mass spectrometry no 5fC signal was observed ($n=3$). This lack of detectable 5fC demonstrated efficient reduction and conversion of 5fC to 5hmC in genomic DNA. Furthermore, 90% of the input total genomic DNA was still detectable by absorbance of a dsDNA binding fluorophore (Qubit assay, Invitrogen), and no fragmentation products were observed following reduction (Supplementary Fig. S5).

The combination of the high yield of 5fC to 5hmC reduction and efficient conversion of the resulting 5hmC to C demonstrates that the redBS-Seq method is both robust and accurate. We next applied redBS-Seq in parallel with oxBS-Seq to accurately determine the levels of 5mC, 5hmC and 5fC at CpG sites throughout the mES cell genome (Fig. 3a,b). We chose to use the reduced representation bisulfite sequencing (RRBS-Seq) technique²⁴, which selectively enriches for CpG dinucleotides, where the majority of methylation sites are known to be located in the genome (23.1 M of total CpG sites), by selective MspI enzyme digestion of CCGG sites. This approach allows each CpG to be sequenced in more depth by suppressing CpG-poor regions.

We treated twelve samples of identical mES cell DNA (four for each method) to either the RRBS-Seq, reduced representation reduced bisulfite sequencing (RRredBS-Seq) or reduced representation oxidative bisulfite sequencing (RRoxBS-Seq) methods to generate quadruplicate data sets for each method. Internal control DNA, containing C, 5mC, 5hmC and 5fC in specific locations and multiple contexts (with DNA containing three 5fCs to nine 5fCs), was spiked into all libraries prior to treatment to quantify the conversion levels of all

cytosine bases and rule out sequence biases by measuring how much each base read as a C following treatment with each method (Fig. 3c). Each method efficiently converted the correct cytosine derivatives, as expected from Fig. 3b, with high reproducibility between replicates.

Sequence data was aligned to the reference mouse genome and the modification percentages calculated at CpG sites (i.e. what percentage read as a C, Supplementary Methods). The distribution of sequence reads across the genome (coverage) was uniform between libraries (Supplementary Fig. S6), indicating the absence of any significant bias introduced by redBS-Seq. The Pearson correlation coefficient (r) of percentage modification (sites that read as C) between pairs of libraries was approximately 0.8 (Supplementary Fig. S7), demonstrating that the reproducibility of each technique was high. As expected, the overall percentage of C modification detected in oxBS-Seq (5mC) was lower than in BS-Seq (5mC + 5hmC), which in turn was lower than in redBS-Seq (5mC + 5hmC + 5fC). Also, the overall modification percentage values were reproducible across the four replicates of the different libraries (Supplementary Table 2).

The presence and levels of 5mC at individual sites was determined from oxBS-Seq, while 5hmC (BS-Seq – oxBS-Seq) and 5fC (redBS-Seq – BS-Seq) from their respective subtractions. Analysis of the data across all positive, but not necessarily significant, sites showed that the average levels of 5mC, 5hmC and 5fC were 13.3%, 3.7% and 3.2%, respectively (Fig. 4a, Supplementary Methods). Statistically significant 5mC was ascertained from any individual sites that contained >1% 5mC, which is the C-U conversion error. Statistically significant 5hmC and 5fC at individual base sites were ascertained using a linear model with binominal distribution from BS-Seq and oxBS-Seq or redBS-Seq (Supplementary Fig. S8, see Supplementary Methods for detailed description). Following determination of significant sites and filtering out sites that were covered by less than 5 reads in 4 or more genomic libraries, we identified 1,015,830 significant 5mC sites averaging 25.4% methylation, 41,013 significant 5hmC sites averaging 17.9% hydroxymethylation and 5,614 significant 5fC sites averaging 22.8% formylation (Fig. 4b, Supplementary Methods). These results demonstrate that even though the global levels of 5fC are lower than 5hmC, there are sites in the genome that contain 5fC at comparable levels to 5hmC.

Symmetry of 5mC, 5hmC and 5fC in CpG context

This data allowed us to investigate to what extent the two cytosines on either strand of a CpG site differ in the levels of modification for 5mC, 5hmC and 5fC, i.e. do modification levels show symmetry. It is known that 5mC levels are symmetric across CpG sites²¹ due to the maintenance mechanism of DNA methylation during replication, which allows patterns of DNA methylation to be inherited through cell cycles²⁵. It has previously been suggested that 5hmC is asymmetric at CpG sites²¹. To quantify average strand symmetry, we measured the fractional difference of the levels of each cytosine modification on either strand of all CpGs where significant modification was detected at both Cs (0= no difference in levels, 1= modification on only one C with complete absence on the other; see Supplementary Methods, Fig. 4c). The asymmetry observed for 5mC is 0.18 (18%), so the

average level of 5mC on the two cytosines on each strand of a CpG differs by just 18%. The asymmetry calculated for 5hmC and 5fC on the other hand was 0.43 (43%) and 0.46 (46%), respectively. This demonstrates that on average for 5hmC and 5fC, one strand contains 43-46% of the levels to the opposite strand, across a population of cells. Interestingly, even at these sites that contain highly asymmetric 5hmC and 5fC, 5mC still has a low asymmetry of 0.2 (20%), indicating that the presence of 5hmC and 5fC has little effect on the symmetry of 5mC. The asymmetry of 5hmC and 5fC, but not 5mC, at CpGs may have fundamental implications for the mechanism of synthesis and/or removal of these bases.

Relationship of 5mC, 5hmC and 5fC genomic levels

It has been proposed that 5mC demethylation might occur through 5hmC and 5fC as intermediates⁷⁻⁹. It has been previously demonstrated that the highest levels of 5hmC are collocated in genomic regions with intermediate levels of 5mC (25-75%) that are suggested to be undergoing a transition in their methylation status¹⁹. Our data suggests that on average the levels of both 5hmC and 5fC are maximal at intermediate levels of 5mC (Fig. 4d,e), consistent with a dynamic pathway by which 5mC levels can be adjusted by being metabolised to 5hmC and 5fC. We also found that in locations where both 5hmC and 5fC were present there was an inverse correlation between the levels of 5fC and 5hmC (Fig. 4f).

It was possible to identify regions in the genome that contained significant levels of 5fC equivalently high or higher than 5mC and/or 5hmC. We have highlighted this in three selected regions containing 5mC, 5hmC and/or 5fC in Fig. 5. These traces demonstrate that, even though global levels of 5fC are low, the levels of 5fC at specific genomic loci can be comparable to the other cytosine modifications.

Discussion

During the preparation of this manuscript, the reduction to practice of an elegant, alternative method, 5fC chemically assisted bisulfite sequencing (fCAB-Seq)¹⁴, was published that can detect 5fC quantitatively and at single-base resolution. In fCAB-Seq, a substituted hydroxylamine is first reacted with 5fC to form a derivative that does not convert to U during bisulfite treatment. Thus, upon comparative analysis of a fCAB-Seq run to a BS-Seq run, the only difference should be 5fC. The conversion efficiency of fCAB-Seq is in the range of 50-60% of 5fC reading as a C following treatment (This range was obtained by extrapolation of data of “% 5fC read as C” vs “% 5fC 76mer Dilution” in Figure S6D of Ref 14). However, here we have demonstrated a 95% conversion efficiency of 5fC reading as a C following redBS-Seq. In both approaches the 5fC to U conversion under bisulfite only conditions is comparable (85-90%, Fig. 2e, Fig 3c and Ref 14). Therefore, it would appear that, redBS-Seq leads to a higher overall sensitivity in 5fC detection and may have advantages in this regard.

In summary, redBS-Seq is a practical method to map 5fC in genomic DNA with precision. The aqueous sodium borohydride reduction of genomic DNA shows a quantitative conversion of 5fC to 5hmC with no detectable fragmentation or side-reactions. The redBS-Seq methodology was applied to the genome of a mouse embryonic stem cell line together with BS-Seq and oxBS-Seq, to generate the first map of 5mC, 5hmC and 5fC at single base

resolution, which provided some new insights into the relationship between 5mC, 5hmC and 5fC in genomic DNA. This practical chemical approach can easily be used to sequence 5fC on a wide variety of platforms including Sanger sequencing and all current next generation platforms. We anticipate that this method will enable the detailed elucidation of the roles of 5fC in various biological contexts in due course.

Methods

General redBS-Seq method

DNA (for quantities see Supplementary Methods) was made up to 15 μ L in water. An aqueous sodium borohydride solution (1 M) was made fresh before every reduction and 5 μ L added to the DNA. The reaction was vortexed and centrifuged, then held at room temperature in the dark for 1 hr. The lids were opened to release any pressure and the reactions vortexed then centrifuged every 15 mins, to remove the bubbles from gas generation. After completion, a sodium acetate solution (10 μ L, 750 mM, pH 5) was added slowly to quench the reaction, violent release of hydrogen gas occurred upon addition. The reaction was then held at room temperature for 10 mins or until no further gas was released.

Reduced samples were subjected to bisulfite treatment using Qiagen Epitect Bisulfite kit with the following alterations; DNA (30 μ L) was combined with bisulfite mix (80 μ L) and DNA protect buffer (30 μ L). The reactions were subjected to two cycles of the FFPE thermal cycle, and worked up as per the manufacturer's protocol for FFPE samples.

Detailed experimental methods are available in the Supplementary Methods

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

We thank the BBSRC for a studentship to M.J.B. and Cancer Research UK for a studentship to M.B.. S.B. is a Senior Investigator of The Wellcome Trust and the Balasubramanian group is core-funded by Cancer Research UK.

M.J.B. and S.B. are co-inventors on a published U.S. patent for redBS-Seq and oxBS-Seq (publication number WO/2013/017853).

References

1. Ndlovu MN, Denis H, Fuks F. Exposing the DNA methylome iceberg. *Trends Biochem. Sci.* 2011; 36:381–7. [PubMed: 21497094]
2. 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]
3. 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]
4. Pfaffeneder T, et al. The discovery of 5-formylcytosine in embryonic stem cell DNA. *Angew. Chem. Int. Ed. Engl.* 2011; 50:7008–12. [PubMed: 21721093]
5. Ito S, et al. Tet proteins can convert 5-methylcytosine to 5-formylcytosine and 5-carboxylcytosine. *Science*. 2011; 333:1300–3. [PubMed: 21778364]

6. Liu S, et al. Quantitative assessment of Tet-induced oxidation products of 5-methylcytosine in cellular and tissue DNA. *Nucleic Acids Res.* 2013; 41:6421–9. [PubMed: 23658232]
7. Hashimoto H, Hong S, Bhagwat AS, Zhang X, Cheng X. Excision of 5-hydroxymethyluracil and 5-carboxylcytosine by the thymine DNA glycosylase domain: its structural basis and implications for active DNA demethylation. *Nucleic Acids Res.* 2012; 40:10203–14. [PubMed: 22962365]
8. 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; 286:35334–8. [PubMed: 21862836]
9. Shen L, et al. Genome-wide analysis reveals TET- and TDG-dependent 5-methylcytosine oxidation dynamics. *Cell.* 2013; 153:692–706. [PubMed: 23602152]
10. 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; 151:1417–30. [PubMed: 23260135]
11. Spruijt CG, et al. Dynamic readers for 5-(hydroxy)methylcytosine and its oxidized derivatives. *Cell.* 2013; 152:1146–59. [PubMed: 23434322]
12. Shen L, et al. Genome-wide analysis reveals TET- and TDG-dependent 5-methylcytosine oxidation dynamics. *Cell.* 2013; 153:692–706. [PubMed: 23602152]
13. Raiber EA, 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:R69. [PubMed: 22902005]
14. Song CX, et al. Genome-wide profiling of 5-formylcytosine reveals its roles in epigenetic priming. *Cell.* 2013; 153:678–91. [PubMed: 23602153]
15. Hayatsu H, Wataya Y, Kai K, Iida S. Reaction of sodium bisulfite with uracil, cytosine, and their derivatives. *Biochemistry.* 1970; 9:2858–65. [PubMed: 5459538]
16. Shapiro R, Servis RE, Welcher M. Reactions of uracil and cytosine derivatives with sodium bisulfite. A specific deamination method. *JACS.* 1970; 92:422–8.
17. Frommer M, et al. A genomic sequencing protocol that yields a positive display of 5-methylcytosine residues in individual DNA strands. *Proc. Natl. Acad. Sci. U.S.A.* 1992; 89:1827–31. [PubMed: 1542678]
18. Huang Y, et al. The behaviour of 5-hydroxymethylcytosine in bisulfite sequencing. *PLoS One.* 2010; 5:e8888. [PubMed: 20126651]
19. Booth MJ, et al. Quantitative sequencing of 5-methylcytosine and 5-hydroxymethylcytosine at single-base resolution. *Science.* 2012; 336:934–7. [PubMed: 22539555]
20. Booth MJ, et al. Oxidative bisulfite sequencing of 5-methylcytosine and 5-hydroxymethylcytosine. *Nat. Protoc.* 2013; 8:1841–51. [PubMed: 24008380]
21. Yu M, et al. Base-resolution analysis of 5-hydroxymethylcytosine in the mammalian genome. *Cell.* 2012; 149:1368–80. [PubMed: 22608086]
22. Oakeley EJ, Schmitt F, Jost JP. Quantification of 5-methylcytosine in DNA by the chloroacetaldehyde reaction. *Biotechniques.* 1999; 27:744–6. 748–50, 752. [PubMed: 10524317]
23. Yu M, et al. Tet-assisted bisulfite sequencing of 5-hydroxymethylcytosine. *Nat. Protoc.* 2012; 7:2159–70. [PubMed: 23196972]
24. Meissner A, et al. Genome-scale DNA methylation maps of pluripotent and differentiated cells. *Nature.* 2008; 454:766–70. [PubMed: 18600261]
25. Bird AP. Use of restriction enzymes to study eukaryotic DNA methylation: II. The symmetry of methylated sites supports semiconservative copying of the methylation pattern. *J. Mol. Biol.* 1978; 118:49–60. [PubMed: 625057]

Figure 1. Outline of reduced bisulfite sequencing to map 5fC in DNA.

a, A specific reduction can convert 5fC to 5hmC and by comparing bisulfite treated reduced and non-reduced DNA, 5fC can be elucidated. **b**, 5fC is converted to uracil under bisulfite treatment. However, upon reduction of 5fC to 5hmC, this base is now converted to cytosine-5-methylsulfonate (CMS) under bisulfite treatment.

Figure 2. Validation of redBS-Seq methodology on synthetic DNA.

5fC can be selectively reduced to 5hmC by sodium borohydride in ssDNA as measured by HPLC analysis, **a**, and dsDNA as measured by mass spectrometry, **b**. Using BS-Seq, 5fC reads as T, **c**, however by employing redBS-Seq, 5fC reads as C, **d**, in 100mer dsDNA with Sanger sequencing. The conversion efficiency of 5hmC and 5fC in a 100mer dsDNA with BS-Seq and redBS-Seq was quantified with Illumina sequencing, **e**, where 5fC efficiency reads as C in redBS-Seq but as T in BS-Seq, with little change in the conversion of 5hmC. Data are the mean and error bars are $\pm$ standard deviation.

Figure 3. Outline of BS-Seq, oxBS-Seq and redBS-Seq and internal control DNA conversion rates.

Outline, **a**, and decoding matrix, **b**, of how the presence of a cytosine modification can be obtained from redBS-Seq, BS-Seq and oxBS-Seq. 5fC and 5hmC are computed by subtracting the C level read in BS-Seq from redBS-Seq and in oxBS-Seq from BS-Seq, respectively. **c**, Analysis of internal control DNA from reduced representative libraries. The percentage of each modified base (C, 5mC, 5hmC and 5fC) in the internal control DNA that reads as a C after each treatment was measured following Illumina sequencing. 5hmC reads as a C in BS-Seq and redBS-Seq but a T in oxBS-Seq. 5fC reads as a T in BS-Seq and oxBS-Seq but a C in redBS-Seq. 5mC reads as a C in all methods and C reads as a T in all methods. Data are the mean and error bars are $\pm$ standard deviation.

Figure 4. Quantification, asymmetry and interplay of 5mC, 5hmC and 5fC in mES cells. 5mC can be directly read from oxBS-Seq. 5hmC can be computed from the difference between BS-Seq and oxBS-Seq. 5fC can be computed from the difference between redBS-Seq and BS-Seq. **a**, Modification levels at CpG sites with positive estimated percentage. **b**, Modification levels of CpG sites with significant levels, as determined by the statistical model (see Supplementary Methods). **c**, Asymmetry of modification levels at CpG sites for all 5mC sites (5mC all), 5mC sites with significant 5hmC or 5fC (5mC with 5hmC/5fC), significant 5hmC sites (5hmC sig.) and significant 5fC sites (5fC sig.). All values are calculated for sites with significant levels of the respective modification on both strands (see Supplementary Methods). **d**, Collocation of 5hmC levels (y-axis) with 5mC (x-axis). **e**, Collocation of 5fC levels (y-axis) with 5mC (x-axis). **f**, Collocation of 5fC levels (y-axis) with 5hmC (x-axis). For all boxplots, data is binned as shown in x-axis labels and the respective percentage ranges associated with each bin are reported on the axis. All boxes span from the 25th to the 75th percentile with median marked by a solid bar. All whiskers extend from the 5th to the 95th percentile. In **a** and **b** outlying data points are also plotted as circles. Colour legend: green is 5mC, red is 5hmC, purple is 5fC.

Figure 5. Visual representation of 5mC, 5hmC and 5fC contexts at three genomic locations. Genomic tracks of regions with clusters of CpG sites, each bar represents levels of 5mC, 5hmC or 5fC at single base resolution. The title to each region consists of the chromosome number (17 (**a**), 1 (**b**) and 14 (**c**)) and the number of base pairs along each chromosome the regions represent. The three regions contain levels of 5fC at levels equivalently high or higher than 5mC and/or 5hmC. Region **a** is located in the PRKD3 gene (protein kinase D3). Region **b** is located between the GM4788 gene (50 kb downstream) and the CFHR1 gene (80 kb upstream). Region **c** is located at the 3' untranslated region (UTR) of the FEZF2 gene (FEZ family zinc finger 2). All regions are 150 base pairs wide.