5-Formylcytosine can be a stable DNA modification in mammals

Martin Bachman[1,2], Santiago Uribe-Lewis[2], Xiaoping Yang[2], Heather E Burgess[3], Mario Iurlaro[3], Wolf Reik[3,4], Adele Murrell[2,5] & Shankar Balasubramanian[1,2]*

Affiliations

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

[2]Cancer Research UK Cambridge Institute, University of Cambridge, Cambridge CB2 0RE, UK

[3]Babraham Institute, Babraham CB22 3AT, UK

[4]Wellcome Trust Sanger Institute, Hinxton, CB10 1SA, UK

[5]Present address: Centre for Regenerative Medicine, Department of Biology and Biochemistry, University of Bath, Bath BA2 7AY, UK

*e-mail: sb10031@cam.ac.uk

Abstract

5-formylcytosine (5fC) is a rare base found in mammalian DNA and thought to be involved in active DNA demethylation. Here, we show that developmental dynamics of 5fC levels in mouse DNA differ from those of 5-hydroxymethylcytosine (5hmC), and using stable isotope labelling *in vivo*, we show that 5fC can be a stable DNA modification. These results suggest 5fC has functional roles in DNA that go beyond being a demethylation intermediate.

DNA of all mammalian cells and tissues is methylated at specific loci, mainly in the 5'-cytosine-phosphate-guanine-3' (CpG) context, to modulate the expression of genes[1]. 5-methylcytosine (5mC) is produced from cytosine (C) by dedicated DNA methyltransferases using S-adenosylmethionine (SAM) as a source of the methyl group[2]. In 2009, two independent laboratories found 5-hydroxymethylcytosine (5hmC) to be present in mammalian DNA and to be the product of ten eleven translocation (TET)-enzyme mediated oxidation of 5mC[3,4]. This oxidised base occurs in all mammalian cells and tissues with global levels ranging between 0.005% and 0.7% of all cytosines[5,6]. The iron(II) and 2-oxoglutarate dependent TET enzymes can also oxidise 5hmC further to 5fC and 5-carboxycytosine (5caC), which were found at levels below 0.002% (or 20 ppm) of all Cs in the genomic DNA of mouse embryonic stem (mES) cells and several adult mouse tissues[7-9].

One proposed role for these oxidised cytosine bases is to serve as intermediates of enzyme-mediated DNA demethylation initiated by oxidation of 5mC[10,11]. Indeed, thymine-DNA glycosylase (TDG) can selectively recognise and excise 5fC and 5caC from the genome and trigger a repair process, which can lead to restoring unmodified C[7,12]. Moreover, mES cells lacking TDG show increased levels of 5fC and 5caC, suggesting that a part of these modifications is constantly being removed from the genome of mES cells[7,12]. On the other hand, we recently demonstrated that 5hmC is a predominantly stable modification in mammalian DNA, especially in the adult mouse brain where 5hmC is most abundant[6]. Herein we investigate the temporal dynamics of 5fC in genomic DNA *in vivo* to consider whether this rare modification can be stable, rather than an active demethylation intermediate (**Fig. 1a**).

We first analysed global levels of all cytosine modifications in the genomic DNA of C57BL/6 mouse tissues to see whether we can detect and quantify 5fC, and identify a relationship between its levels and those of its precursors 5mC and 5hmC or its metabolite 5caC. We included a range of postnatal tissues from newborn (1 d old), adolescent (21 d old) and adult (15 w old) mice, as their genomic DNA is known to have different levels of 5hmC depending on the overall proliferation rate (and therefore the age) of the tissue[6]. We also included C57BL/6 embryos at 11.5 d post-fertilisation as this is the lethal age for mice lacking TDG[13,14], and mES cells derived from the same strain were added for comparison. To achieve quantification of the rare modifications (5fC and 5caC) with the highest possible sensitivity and accuracy, we employed a nano high-performance liquid chromatography – tandem high-resolution mass spectrometry (nanoHPLC-MS/HRMS) method, which is able to resolve genuine rare modified bases (5fC and 5caC) from potential impurities of the same nominal mass and retention time, and can detect down to 0.1 ppm of total cytosines in as little as 100 ng of digested genomic DNA. In addition, the use of isotopically labelled internal standards (IS) of C, 5mC and 5hmC substantially improved the quality of the measurements, ensured excellent reproducibility between technical replicates and excluded spontaneous oxidation of 5hmC as the source of 5fC or 5caC. Example mass spectra, extracted ion chromatograms and calibration curves are shown in **Supplementary Results**, **Supplementary Figs. 1-5**. 5mC and

5hmC were present in all tissues (**Supplementary Fig. 6**) and their levels were in good agreement with available published data[5,6,8,9]. While 5mC levels show a relatively uniform distribution between tissues, global 5hmC content is highly correlated to the proportion of proliferating cells in the tissue as we have shown previously[6]. We found 5fC to be also present in all studied tissues at levels ranging between 0.2 ppm and 15 ppm of all cytosines (**Fig. 1b** and **Supplementary Fig. 6**). Notably, 5caC was not detected in any postnatal tissues from C57BL/6 mice, even in those with high 5fC content, but several tissues from C57BL/6 embryos (**Fig. 1b**) and adult (12 w old) CD1 mice (**Supplementary Fig. 7**) contained up to 2 ppm of this rare DNA base modification. Overall, we found no correlation between the levels of 5fC and the levels of its precursors 5mC or 5hmC (**Supplementary Fig. 8**), nor did we find any clear pattern of DNA modification changes as the tissues age. They can retain the levels of 5fC while gaining 5hmC (e.g. brain), lose 5fC while retaining the levels of 5hmC (e.g. heart), or even lose 5fC while gaining 5hmC (e.g. liver) (**Figs. 1c-e**). We found that DNA from mES cells lacking all three TET enzymes (TET triple-knockout (TET-TKO))[15] contains no detectable 5hmC, 5fC or 5caC (**Fig. 1a** and **Supplementary Fig. 6**), confirming that 5hmC is the only source of 5fC and 5caC in mES cell DNA. Although we have no measure of tissue-specific susceptibility to oxidation (such as the quantity of the oxidative lesion 8-oxoguanine), the lack of correlation between global levels of 5hmC and 5fC (**Supplementary Fig. 8**) together with the lack of positional overlap between 5hmC and 5fC in mES cells[16,17] strongly suggest that 5fC and 5caC are not generated by spontaneous oxidation of 5hmC and 5fC.

To elucidate the stability of 5fC in genomic DNA towards turnover *in vivo*, we applied a stable isotope tracing method consisting of feeding cultured cells and mice with [*methyl*-$^{13}CD_3$] L-methionine as we have done previously to study the lifetime of 5hmC[6]. The *methyl*-$^{13}CD_3$ group enters the intracellular pool of SAM and is transferred into newly methylated cytosines by the action of DNMT enzymes. The TET enzymes can then convert labelled [*methyl*-$^{13}CD_3$] 5mC (5mC[+4]) into [*hydroxymethyl*-$^{13}CD_2$] 5hmC (5hmC[+3]) and [*formyl*-^{13}CD] 5fC (5fC[+2]) (**Fig. 1a**). The labelling ratios (e.g. % 5fC[+2] over total 5fC) change according to the dynamics and half-life of the given modification in the genomic DNA. For example, a modification that is quickly turning over in DNA would show a high labelling ratio, whereas a very stable modification would show no labelling in non-proliferating cells or tissues. The maximum obtainable ratio also depends on the activity of other biosynthetic pathways feeding into the one carbon metabolism. The labelling ratios can be determined very accurately for each modified cytosine using LC-MS/HRMS due to unique masses of the labelled base fragments (**Supplementary Fig. 8**). We first cultured mES cells in the labelled ([*methyl*-$^{13}CD_3$] L-methionine) media for 8 days, and found that the labelling ratio of 5fC increases much slower than that of 5mC and 5hmC. This indicates either a substantial time lag in making 5fC from newly formed 5hmC, or presence of a population of slower or non-dividing (unlabelled) cells with higher global levels of 5fC compared to the fast dividing (labelled) population of mES cells (**Fig. 2a**).

We then analysed genomic DNA from C57BL/6 mice fed with a diet where all L-methionine was replaced with [*methyl*-$^{13}CD_3$] L-methionine. To gain information about 5fC in developing tissues, we fed a pregnant female starting from 7 d before birth, and kept the family for 6 more days on the labelled diet (the 6 d-old pups were therefore labelled for 13 d when harvested). The genomic DNA in tissues such as kidney or colon showed uniform labelling of around 30% for all detectable modifications (5mC, 5hmC and 5fC) (**Fig. 2b**). However, brain tissue from the same pups showed much less 5hmC[+3] and no detectable 5fC[+2]. This indicates that 5fC was formed in these tissues prior to the start of labelling and remained there for 13 d until the DNA was harvested. 1 d-old newborns labelled from conception and with pre-labelled parents (52 d prior to conception, total labelling time of pup is therefore 22 d) already showed a higher 5fC labelling in the brain, but the ratio was still lower than those of 5mC and 5hmC (27% vs. 44% and 42.5%, respectively) (**Fig. 2b**). During the gestation period, the labelling ratio of the methionine pool in the pregnant female was still increasing and therefore this observation is consistent with 5fC being more abundant on the older or slower proliferating DNA as concluded above. Proliferating tissues from adult mice (e.g. spleen) showed a similar trend where the 5fC labelling ratio was always smaller than that of 5mC or 5hmC, even in animals labelled for as long as 4 months (117 d) (**Fig. 2c**). This effect is best explained by 5fC being mostly stable, and again by the presence of non-dividing (unlabelled) cells alongside a population of proliferating (labelled) cells that have lower global 5fC levels than the non-dividing cells.

Finally, in the mostly non-dividing adult brain where only 1.3% and 3.7% of 5mC becomes labelled during the 117-d feeding period, there was no detectable labelled 5fC. Again, this is not due to a lack of an intracellular pool of [*methyl*-$^{13}CD_3$] SAM, as we could measure more than 50% [*methyl*-$^{13}CD_3$]

5mC in RNA in adult brain and cerebellum (**Fig. 2d**). If 5fC was a short-lived DNA modification and was constantly being turned over, its labelling ratio would be close to the labelling ratio of intracellular SAM and 5mC in RNA. If 5fC was short-lived and only produced from pre-existing unlabelled 5hmC in the adult brain, the levels of 5hmC would be depleted over time, which is not consistent with the high levels of 5hmC in this tissue and is the opposite of what has been described for ageing brain[18]. Therefore, the lack of 5fC labelling in the adult brain means that this modified base must be stable in the genome as opposed to generally acting as a dynamic intermediate of active DNA demethylation.

In summary, we present the first direct evidence that 5fC (derived from mC by TET-mediated oxidation) can be a stable DNA modification *in vivo*, and provide quantitative measurements of the levels of all modified cytosines in mouse tissues across several developmental stages. 5fC levels do not correlate with those of its precursors 5mC and 5hmC, its metabolite 5caC or with age of the individual. Whilst there is precedent for removal of 5fC and 5caC from the genome (e.g. in mES cells), probably in the process of active DNA demethylation[7,12], our findings suggest that the bulk of 5fC can be stable. 5fC has been identified as having more protein binders than 5mC or 5hmC[19,20] and having a distinct genomic profile from 5mC, 5hmC or 5caC at single-base resolution[12,16,17,21-23]. Moreover, 5fC has recently been shown to alter the structure of the DNA double helix[24]. We therefore conclude that such stably 5fC-modified DNA could have profound consequences for the regulation of gene expression that may be distinct to those caused by the presence of 5mC and 5hmC. However, direct evidence regarding the biological function of 5fC remains to be demonstrated.

Methods

Methods and any associated references are available in the online version of the paper.

References

1. Ooi, S.K., O'Donnell, A.H. & Bestor, T.H. Mammalian cytosine methylation at a glance. *J Cell Sci* **122**, 2787-91 (2009).
2. Goll, M.G. & Bestor, T.H. Eukaryotic cytosine methyltransferases. *Annu Rev Biochem* **74**, 481-514 (2005).
3. Kriaucionis, S. & Heintz, N. The nuclear DNA base 5-hydroxymethylcytosine is present in Purkinje neurons and the brain. *Science* **324**, 929-30 (2009).
4. Tahiliani, M. et al. Conversion of 5-methylcytosine to 5-hydroxymethylcytosine in mammalian DNA by MLL partner TET1. *Science* **324**, 930-5 (2009).
5. Globisch, D. et al. Tissue distribution of 5-hydroxymethylcytosine and search for active demethylation intermediates. *PLoS One* **5**, e15367 (2010).
6. Bachman, M. et al. 5-Hydroxymethylcytosine is a predominantly stable DNA modification. *Nat Chem* **6**, 1049-55 (2014).
7. He, Y.F. et al. Tet-mediated formation of 5-carboxylcytosine and its excision by TDG in mammalian DNA. *Science* **333**, 1303-7 (2011).

8. Ito, S. et al. Tet proteins can convert 5-methylcytosine to 5-formylcytosine and 5-carboxylcytosine. *Science* **333**, 1300-3 (2011).

9. Pfaffeneder, T. et al. Tet oxidizes thymine to 5-hydroxymethyluracil in mouse embryonic stem cell DNA. *Nat Chem Biol* **10**, 574-81 (2014).

10. Song, C.X. & He, C. Potential functional roles of DNA demethylation intermediates. *Trends Biochem Sci* **38**, 480-4 (2013).

11. Schiesser, S. et al. Deamination, oxidation, and C-C bond cleavage reactivity of 5-hydroxymethylcytosine, 5-formylcytosine, and 5-carboxycytosine. *J Am Chem Soc* **135**, 14593-9 (2013).

12. Neri, F. et al. Single-Base Resolution Analysis of 5-Formyl and 5-Carboxyl Cytosine Reveals Promoter DNA Methylation Dynamics. *Cell Reports* **10**, 674-683 (2015).

13. Cortazar, D. et al. Embryonic lethal phenotype reveals a function of TDG in maintaining epigenetic stability. *Nature* **470**, 419-23 (2011).

14. Cortellino, S. et al. Thymine DNA glycosylase is essential for active DNA demethylation by linked deamination-base excision repair. *Cell* **146**, 67-79 (2011).

15. Hu, X. et al. Tet and TDG mediate DNA demethylation essential for mesenchymal-to-epithelial transition in somatic cell reprogramming. *Cell Stem Cell* **14**, 512-22 (2014).

16. 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**, 435-40 (2014).

17. Song, C.X. et al. Genome-wide profiling of 5-formylcytosine reveals its roles in epigenetic priming. *Cell* **153**, 678-91 (2013).

18. Kraus, T.F., Guibourt, V. & Kretzschmar, H.A. 5-Hydroxymethylcytosine, the "Sixth Base", during brain development and ageing. *J Neural Transm*, doi:10.1007/s00702-014-1346-4 (2014).

19. Iurlaro, M. et al. A screen for hydroxymethylcytosine and formylcytosine binding proteins suggests functions in transcription and chromatin regulation. *Genome Biol* **14**, R119 (2013).

20. Spruijt, C.G. et al. Dynamic readers for 5-(hydroxy)methylcytosine and its oxidized derivatives. *Cell* **152**, 1146-59 (2013).

21. Booth, M.J. et al. Quantitative sequencing of 5-methylcytosine and 5-hydroxymethylcytosine at single-base resolution. *Science* **336**, 934-7 (2012).

22. Yu, M. et al. Base-resolution analysis of 5-hydroxymethylcytosine in the mammalian genome. *Cell* **149**, 1368-80 (2012).

23. Lu, X. et al. Base-resolution maps of 5-formylcytosine and 5-carboxylcytosine reveal genome-wide DNA demethylation dynamics. *Cell Res* **25**, 386-9 (2015).

24. Raiber, E.A. et al. 5-Formylcytosine alters the structure of the DNA double helix. *Nat Struct Mol Biol* **22**, 44-9 (2015).

25. Ying, Q.L. et al. The ground state of embryonic stem cell self-renewal. *Nature* **453**, 519-23 (2008).

Acknowledgments

We would like to thank Clive d'Santos and David Oxley for their support with mass spectrometry, and Guoliang Xu for kindly providing TET-TKO mES cells. This work was supported by the Cancer Research UK (C14303/A17197, S.B.), The Wellcome Trust (WT099232, S.B.; WT095645/Z/11/Z, W.R.) and the BBSRC (BB/K010867/1, W.R.).

Author contributions

M.B., S.U.L. and S.B. conceived the study; S.U.L., M.B., H.B. and M.I. performed experiments; M.B. and X.Y. carried out mass spectrometry and data analysis; S.B., A.M and W.R. supervised the project; M.B. and S.B. wrote the manuscript with contributions from all authors.

Competing financial interests

S.B. and W.R. are founders and advisors to Cambridge Epigenetix, Ltd.

Additional information

Supplementary Figures 1-9 and Supplementary Table 1. Correspondence and requests for materials should be addressed to S.B.

Online Methods

Animals. All *in vivo* experiments were performed under the terms of a UK Home Office license. C57BL/6 and CD1 mice were bred and housed according to UK Home Office guidelines. Custom L-methionine-free mouse diet supplemented with [*methyl*-$^{13}CD_3$] L-methionine (Sigma) was manufactured by TestDiet.

Cell culture. mES cells were derived by Dr Xiangang Zou in the CRUK Cambridge Institute from a C57BL/6 mouse (Charles River) and cultured on a gelatin-coated plate in a DMEM-KO medium (Invitrogen) supplemented with

10% FCS, MEM non-essential amino acids, glutamine, sodium pyruvate, penicillin, streptomycin, mouse leukemia inhibitory factor (mLIF) and 2i as described by Ying *et al*.[25] TET-TKO mES cells were obtained from Guoliang Xu[15] and cultured in the 2i conditions as above. All cells were regularly tested for mycoplasma contamination. For isotopic labelling experiments, cells were maintained in a custom L-methionine-free DMEM-KO medium (Invitrogen) supplemented with 30 mg/L of [*methyl*-$^{13}CD_3$] L-methionine (Cambridge Isotope), and the respective components above.

Genomic DNA extraction. Tissues and cells were resuspended in lysis buffer (100 mM Tris, pH 5.5, 5 mM EDTA, 200 mM NaCl, 0.2% SDS) supplemented with 400 µg/ml proteinase K (Invitrogen), and were incubated at 55°C overnight. DNA was purified using phenol:chloroform:isoamyl alcohol (25:24:1, Sigma) and Phase Lock Gel (5 Prime), precipitated from 70% ethanol and resuspended in ultrapure HPLC-grade water.

DNA degradation to 2'-deoxynucleosides and LCMS analysis. 1-2 µg of DNA was incubated with 5 U of DNA Degradase Plus (Zymo Research) in a total volume of 30 µl for 4 h at 37°C. Samples were filtered through a pre-washed Amicon 10 kDa centrifugal filter unit (Millipore) before LCMS analysis.

LCMS analysis of global 5mC, 5hmC, 5fC and 5caC levels. Analysis of global levels of 5mC, 5hmC, 5fC and 5caC was performed on a Q Exactive mass spectrometer (Thermo) fitted with an UltiMate 3000 RSLCnano HPLC (Dionex) and a self-packed hypercarb column (20 mm × 75 µm, 3 µm particle size) at a flow rate of 0.75 µl/min, and a gradient of 0.1% formic acid in water and acetonitrile. Calibration curves were generated using a mixture of synthetic standards 2'-deoxycytidine (Sigma), 5-methyl-, 5-hydroxymethyl-, 5-formyl- and 5-carboxy-2'-deoxycytidine (Berry&Associates), in the ranges of 0.5 nM – 5 µM for C, 0.025 – 250 nM for 5mC and 0.005 – 50 nM for 5hmC, 5fC and 5caC. Samples and synthetic standards were spiked with an isotopically labelled mix containing 100 nM of 2'-deoxycytidine-(^{15}N,d$_2$) (synthesis and characterisation in Bachman *et al*.[6]), 5-methyl-2'-deoxycytidine-(d$_3$) and 5-hydroxymethyl-2'-deoxycytidine-(d$_3$) (both Toronto Research Chemicals). Target ions were fragmented in a positive ion mode at 10% normalized collision energy, and full scans (50 – 300 Da) were acquired. The inclusion list contained the following masses: C (228.1), C_IS (231.1), 5mC (242.1), 5mC_IS (245.1), 5hmC (258.1), 5hmC_IS (261.1), 5fC (256.1), 5caC (272.1). Extracted ion chromatograms of base fragments (see Supplementary Fig. 1) were used for quantification. Results are expressed as a % or ppm of total cytosines.

LCMS analysis of isotope incorporation into genomic DNA*.* Analysis of isotope incorporation into DNA was performed using the same instrumental set up as above, targeting ions of masses 242.1 (mC), 246.1 (mC[+4]), 258.1 (5hmC and 5fC[+2]), 261.1 (5hmC[+3]) and 256.1 (5fC). Extracted ion chromatograms of base fragments were used for quantification of labelling ratios (see also Supplementary Fig. 8). Results are expressed as % labelling (e.g. % 5fC[+2] represents the percentage of labelled 5fC[+2] in total 5fCs).

LCMS analysis of isotope incorporation into 5mC in RNA. Total RNA was carried through during genomic DNA extraction (no RNase treatment), during hydrolysis to nucleosides and LC-MS/HRMS analysis. An additional mass of 262.1 was targeted for RNA 5mC[+4] (unlabelled 5mC was present in the 258.1 channel), and base fragments 126.0662 (5mC) and 130.0884 (5mC[+4]) were used for quantification of % labelling as above.

Figure Legends:

Figure 1 | Dynamics of global levels of 5fC during mouse development are distinct to those of 5hmC. (**a**) Metabolism of cytosine modifications in DNA. While the majority of 5mC and 5hmC persist in the genomic DNA, the stability of 5fC and 5caC *in vivo* was unknown. Feeding [*methyl*-^{13}CD$_3$] L-methionine can be used to measure the lifetime of cytosine modifications in cells and in vivo. Labelling pattern is indicated in red. See also **Fig. 2**. (**b**) Global levels of 5fC and 5caC in genomic DNA from mouse embryos (E11.5). Shown are mean ± SEM of 3 animals. Each sample was analysed in technical duplicate and the mean value was used. (**c, d, e**) Changes of global 5fC, 5mC and 5hmC levels during development in selected C57BL/6 mouse tissues (further data in **Supplementary Fig. 6**). Shown are mean ± SEM of 3 embryos (E11.5) (data from **Fig. 1b**), 3 newborns (1 d old), 2 adolescent (21 d old) and 2 adult (15 w old) mice. See also **Supplementary Figs. 7–8**.

Figure 2 | 5fC can be a stable DNA modification *in vivo*. (**a**) Labelling ratios of 5mC, 5hmC and 5fC in the genomic DNA of mES cells cultured in the presence of [*methyl*-$^{13}CD_3$] L-methionine. Shown are single measurements and total labelling time is given in brackets. (**b, c** and **d**) Labelling ratios of 5mC, 5hmC and 5fC in the genomic DNA of C57BL/6 mice fed with the [*methyl*-$^{13}CD_3$] L-methionine diet. (**b**) 6 d-old pups labelled from 1 week prior to birth (total labelling time of 13 d), or 1 d-old newborn (total labelling time of 22 d, parents on labelled diet for 52 d prior to conception). Shown are mean ± SEM of 2 animals (6 d-old pups) or 2 technical replicates (1 d-old pup). (**c** and **d**) Mice labelled in adulthood. Shown are mean ± SEM of at least 2 technical replicates from individual mice, and total labelling time is shown in brackets. The absence of 5fC[+2] in the brain (**d**) where 5fC is most abundant (see **Fig. 1** and **Supplementary Fig. 6**) indicates minimal or no further generation of 5fC once placed in post-mitotic tissues. Moreover, if 5fC was involved in cycles of methylation and demethylation, its labelling ratio would be similar to that of 5mC in RNA (**d**).