5-Formylcytosine to Cytosine Conversion by C-C Bond Cleavage *in vivo*

Katharina Iwan[1], René Rahimoff[1], Angie Kirchner[1], Fabio Spada[1], Arne S. Schröder, Olesea Kosmatchev, Shqiponja Ferizaj, Jessica Steinbacher, Edris Parsa, Markus Müller, Thomas Carell*

Center for Integrated Protein Science Munich CiPS[M] at the Department of Chemistry, Ludwig-Maximilians-Universität München, 81377 Munich, Germany

[1] These authors contributed equally

*Corresponding author: Thomas Carell, Thomas.carell@lmu.de

Abstract

Tet enzymes oxidise 5-methyl-deoxycytidine (mdC) to 5-hydroxymethyl-dC (hmdC), 5-formyl-dC (fdC) and 5-carboxy-dC (cadC) in DNA. It was proposed that fdC and cadC deformylate and decarboxylate to dC in the course of an active demethylation process. This would re-install canonical dC bases at previously methylated sites. The question whether such direct C-C bond cleavage reactions at fdC and cadC occur *in vivo* remains an unsolved problem. Here we report the incorporation of synthetic isotope- and (*R*)-2'-fluorine-labelled dC and fdC-derivatives into the genome of cultured mammalian cells. Following the fate of these probe molecules using UHPLC-MS/MS provided quantitative data about the formed reaction products. The data show that the labelled fdC probe is efficiently converted into the corresponding labelled dC, most likely after its incorporation into the genome. This allows concluding that fdC is undergoing C-C bond cleavage in stem cells that leads to the direct re-installation of unmodified dC.

Introduction

Modification of genomic cytosine modulates the interaction of DNA-binding factors to the genome, thus affecting gene expression and chromatin structure. The primary and most prevalent modification is methylation to 5-methylcytidine (mdC), which in mammals is catalyzed by the DNA methyltransferases Dnmt1, 3a and 3b, at least partly in co-operation with the catalytically inactive Dnmt3l. Because Dnmt1 is a maintenance methyltransferase that copies the methylation pattern during replication, the information that they convey is inherited through cell division. Genomic mdC can be iteratively oxidised to 5-hydroxymethyldeoxycytidine (hmdC)[1,2], 5-formyldeoxycytidine (fdC)[3,4], and 5-carboxydeoxycytidine (cadC)[4,5] by the Ten-eleven translocation (Tet) family of α-ketoglutarate dependent dioxygenases (Fig. 1a). These oxidised cytidine derivatives are prominently detected in DNA isolated from neuronal tissues[2,6,7] and in the genome of embryonic stem cells (Fig. 1b), where their levels change during differentiation.[1,4,8] For example, hmdC can reach levels of up to 1.3% per dG in DNA isolated from brain.[9] While the presence of mdC and hmdC is believed to influence the transcriptional activity of genes[10,11], no clear function could yet be assigned to the higher oxidised modifications fdC and cadC. Recent reports, however, show that fdC is a stable[12] or at least semi-stable[13] base in the genome. These discoveries and the finding of specific reader proteins that recognise fdC and cadC suggest that they might have regulatory purposes as well.[14-17] So far, however, fdC and cadC are mainly considered to be intermediates of an active demethylation process that allows cells to replace mdC by a canonical dC nucleotide.[17-19] One such scenario involves that fdC and cadC are substrates of the Thymine-DNA glycosylase (Tdg), which cleaves the corresponding glycosidic bond. This converts fdC and cadC into abasic sites, which are further processed through base excision repair (BER) as depicted in figure 1a. This Tdg-initiated process establishes an active demethylation pathway,

which ultimately incorporates unmodified dC nucleotides at former fdC and cadC sites.[5,20] A problem associated with the mechanism is that the removal of every mdC creates a potentially harmful single strand break intermediate. In case another mdC is close to the first in the opposite DNA strand, even double strand breaks may be generated. In addition to these concerns, it was shown that both maternal and paternal genomes of mouse zygotes undergo active demethylation independently of Tdg.[21] To explain this and provide an alternative to the generation of harmful repair intermediates, it was suggested that fdC and cadC may directly deformylate and decarboxylate, respectively, under C-C bond cleavage (Fig. 1a).[6,22] Indeed, chemical pathways that allow such a direct deformylation and decarboxylation of fdC and cadC have been described.[23,24] They involve addition of a helper nucleophile to the C6-position of fdC and cadC in a Michael-addition-type reaction, followed by deformylation or decarboxylation and final elimination of the helper nucleophile.[23] The chemistry is therefore quite similar to the known reaction mechanisms employed by the Dnmt proteins. Although chemically plausible, it is unclear if such direct C-C bond cleavage reactions occur within the genome.[25] This process would provide a new and harmless way to convert mdC back into dC in the genome without the formation of potentially harmful abasic site intermediates.

Here we report a sensitive mass spectrometry-based isotope tracing study to investigate if a C-C bond cleavage reaction occurs on fdC bases (Fig. 1c). We supplemented the medium of cultured mammalian cells with synthetic isotope and fluorine labelled fdC derivatives to metabolically integrate the nucleosides as reporter molecules into their genome. After isolation of the genomic DNA, the levels of the modified dC derivatives were measured by ultra-high pressure liquid chromatography coupled to tandem mass spectrometry (UHPLC-MS/MS), thereby tracing isotopically or fluorine-labelled dC derivatives.
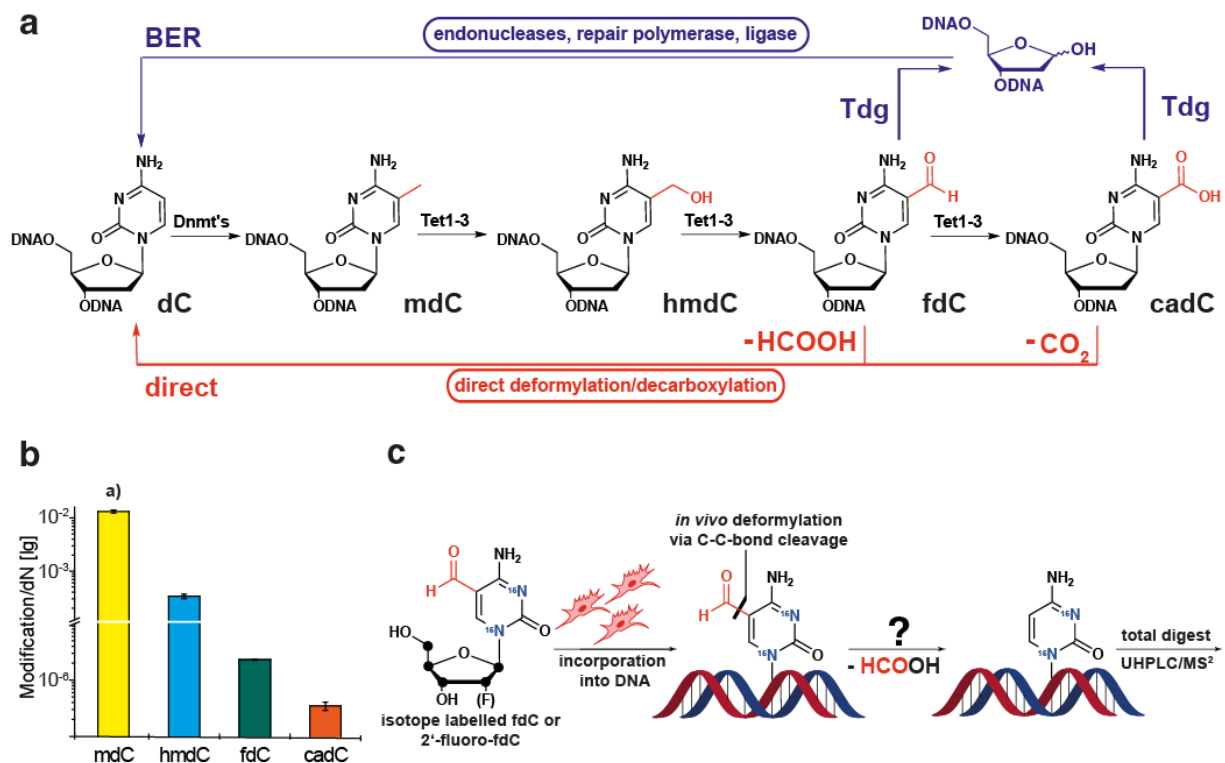
Figure 1 | Isotope tracing studies. **(a)** Suggested pathways of active demethylation. Tdg based cleavage of the glycosidic bond of fdC and cadC results in an abasic site, which initiates a BER process that leads to the replacement of fdC and cadC by canonical dC (blue). Deformylation of fdC (-HCOOH) and decarboxylation (-CO_2) of cadC provides dC directly (red). **(b)** UHPLC-coupled MS/MS experiments allow exact quantification of various dC derivatives in mESCs. Mean and s.d. of technical triplicates from two independent cultures are shown. **(c)** Schematic depiction of the feeding experiment using synthetic isotope and fluorine labelled fdC derivatives, which are metabolically integrated into the genome. ● = ^{13}C atoms.

Results

Isotopically labelled fdC is directly converted into dC in mouse embryonic stem cells

We started the study with a $[^{13}C_5][^{15}N_2]$-fdC **1** (Fig. 2a) in which all five C-atoms of the ribose ring were exchanged by ^{13}C and the two in-ring nitrogen atoms replaced with ^{15}N (Supplementary Note). This provides compound **1**, which is seven mass units heavier than naturally occurring fdC and hence easily distinguishable by mass spectrometry. The large mass difference allows exact tracing of all transformations that may take place on this base with high sensitivity, because the natural abundance of such highly isotopically modified dC-derivatives is essentially null. Possible transformations are the deformylation of **1** to $[^{13}C_5][^{15}N_2]$-dC **2** and its deamination to $[^{13}C_5][^{15}N_2]$-dU **3** followed by methylation of **3** to $[^{13}C_5][^{15}N_2]$-dT **4**. Alternatively, compound **1** can deaminate to $[^{13}C_5][^{15}N_2]$-fdU **5** and finally the deformylated compound **2** could be methylated to $[^{13}C_5][^{15}N_2]$-mdC **6**. Analysis of the MS pattern of **1** shows that cleavage of the glycosidic bond is the dominant fragmentation pathway. This leads to a clearly detectable fingerprint mass transition of $m/z = 263.1 \rightarrow m/z = 142.1$ (Fig. 2b). As an example,

detection of the demodified product dC **2** would be possible based on its mass transition from m/z = 235.1 to m/z = 114.0. For the experiment, we added **1** to the medium of mouse embryonic stem cells (mESCs) under priming conditions. After three days, the genomic DNA was isolated using a standard protocol and digested to the individual nucleosides. The obtained nucleoside mixture was analysed by UHPLC coupled to a triple quadrupole mass spectrometer. We noted that **1** was indeed metabolically incorporated into the genome of mESCs. The mass transition of **1** (m/z = 263.1 → m/z = 142.1) was clearly detectable at a retention time of 5.50 min under our conditions (Fig. 2b). By using the mass transitions specific for all the expected natural dC-derivatives, we were also able to detect next to **1**, mdC, hmdC and fdC (Fig. 1b).

Analysis of the nucleoside mixtures revealed the presence of a new dC-derivative at a retention time of 1.95 min, which showed the expected mass transition (m/z = 235.1 → m/z = 114.0) for **2**, showing that $[^{13}C_5][^{15}N_2]$-fdC is indeed demodified (Fig. 2c). In order to unequivocally prove the identity of **2**, an even heavier isotopically modified dC-isotopologue $[^{13}C_9][^{15}N_3]$-dC **7**, with a characteristic MS transition of m/z = 240.1 → m/z = 119.1 (Supplementary Fig. 1) was used as an internal standard. Compound **7** was added to the nucleoside mixture and it eluted at the same retention time as **2** (Fig. 2c), confirming that the UHPLC-MS/MS signal at 1.95 min is caused by **2**. Exact quantification of the conversion showed that when **1** was supplied to mESC cultures at 50 µM for three days, steady-state incorporation levels of about 3 - 5x10^{-7} of $[^{13}C_5][^{15}N_2]$-fdC per dN in genomic DNA were reached (Fig. 2d). We observed higher levels of product **2** (up to a factor of 10), as shown in figure 2d.

While we can exclude spontaneous deformylation of **1** based on previous stability studies[23], **2** can in principle form either by C-C bond cleavage in the genome or at the level of the soluble nucleoside/nucleotide pool. Conversion of **1** in the soluble pool to **2** would then be followed by metabolic incorporation of the **2**-triphosphate into the genome. It is known that soluble **2** is the substrate for other metabolic processes such as deamination to **3** (catalysed by cytidine deaminase and deoxycytidylate deaminase) followed by methylation by thymidylate synthase to give **4**.[26,27] In order to distinguish the two possible conversion scenarios (genomic DNA vs. soluble pool), we reasoned that if **1** is converted into **2** in the soluble pool, we must find compounds **3** and particularly **4** in the genome.
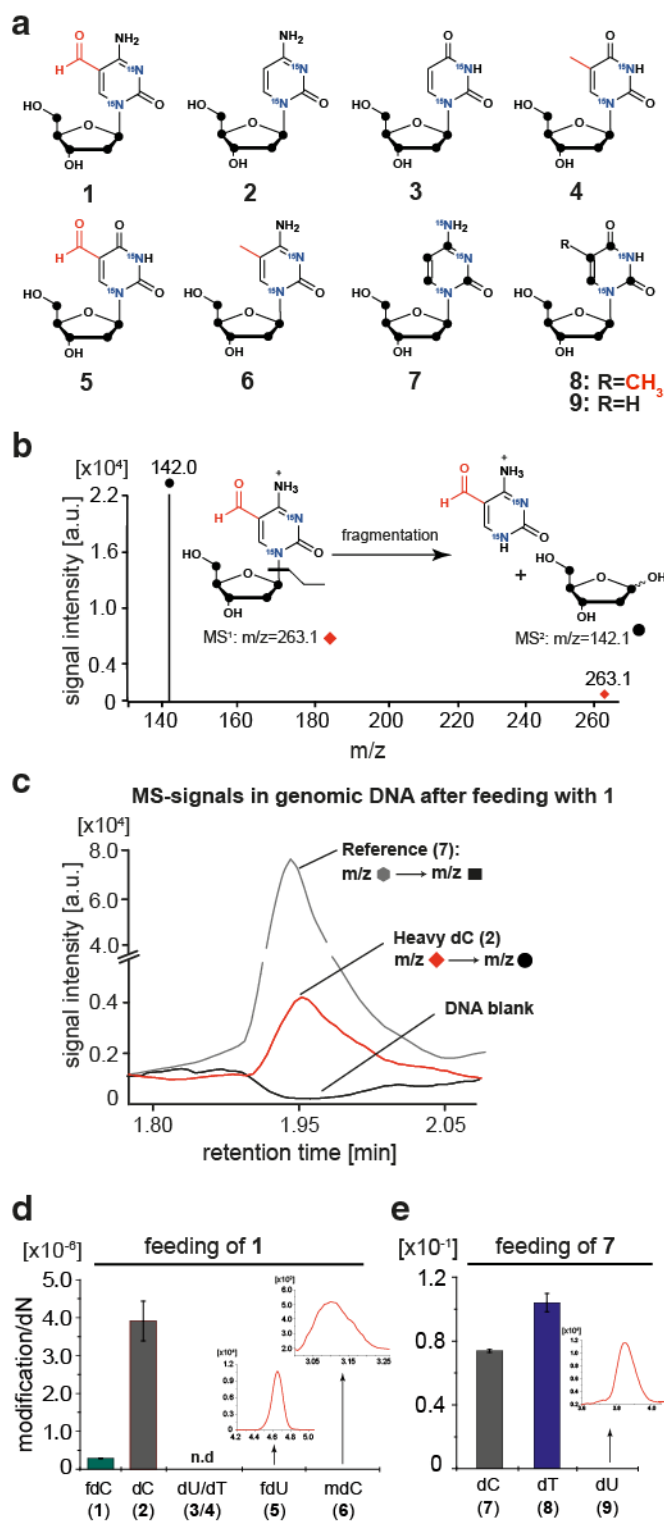
Figure 2 | Conversion of isotopically labelled fdC into dC in mESCs. (a) Overview of the compounds that can be detected after feeding of 1 to mESCs. (b) Feeding of 1 to mESCs results in incorporation of the isotopologue into the genomic DNA as proven by its fingerprint MS transition. (c) Analysis of gDNA after feeding of 1 shows the presence of labelled dC 2. (d) Quantitative data obtained upon feeding 1 to mESCs. Mean and s.d. of technical triplicate measurements from a single culture are shown. (e) Quantitative data obtained upon feeding of 7 to mESCs. Mean and s.d. of technical triplicates from two independent cultures are shown. For a schematic overview of the dC or fdC metabolic pathways see Supplementary Figure 3.

In order to investigate the behaviour of soluble dC, we cultured mESCs in the presence of a isotopically labelled dC-derivative $[^{13}C_9][^{15}N_3]$-dC, and indeed detected the expected presence of the corresponding isotopically labelled deamination products $[^{13}C_9][^{15}N_2]$-dT/dU **8/9** in the genome (Fig. 2e). In contrast, when **1** was supplied to mESC cultures, we detected next to **2** only the direct deamination product **5** in the genome, but not **3** and **4** (Fig. 2d). Even upon feeding of **1** to mESCs for three consecutive days, we were unable to detect even traces of **4** in the genome. This argues against formation of **2** in the soluble pool. We next analysed the soluble nucleoside/nucleotide pool directly for the content of **2**, after feeding of **1** (Supplementary Fig. 2). To this end, we fed **1** to mESCs over three days. The cells were washed extensively and finally resuspended in 50% (v/v) MeCN to extract soluble metabolites. After further purification by solid-phase extraction, the nucleotides were dephosphorylated to nucleosides. Analysis of this solution by UHPLC-MS/MS did not give any signal for **2**. All these control experiments suggest that **1** undergoes C-C bond cleavage to **2** directly in the genome and not in the soluble pool,

although such a scenario cannot be fully ruled out due to the complexity of the metabolic pathways (Supplementary Fig. 3). Interestingly, we also noted the presence of the re-methylated product **6** in the genome of mESCs fed with **1**, but due to the low signal intensity, we were unable to obtain quantitative data (Fig. 2d).

(*R*)-2'-fluorinated cytosines are useful probes for biochemical conversions *in vivo*

In order to reach higher sensitivity, we experimented with various other isotope labelled fdC derivatives and finally found that 2'-fluorinated dC derivatives **10** - **17** are excellent probe molecules (Fig. 3a and Supplementary Fig. 4). The F-atom makes the compounds 18 atom units heavier. The compounds have a slightly shifted retention time (Fig. 3b) and give sharp signals in the UHPLC-MS/MS analysis because of a glycosidic bond that is more labile in the MS-fragmentation step. Furthermore, the 2'-(*R*)-configured compounds are well tolerated by the cells used for this study. The F-substituent does affect the ability of the molecule to undergo further biochemical conversions, but the effect is small. (*R*)-2'-F-dC **10** is for example efficiently methylated by DNA methyltransferases[28] and (*R*)-2'-F-mdC **11** is also oxidised to (*R*)-2'-F-hmdC **12** by the Tet enzymes (Fig. 3a+c), although here a reduced speed of oxidation is observed.[29]

To show that the fluorinated compounds are valid probe molecules, we first added **10** to the mESC culture at 0.5 µM, 1.0 µM or 2.5 µM for three days. Under these conditions, UHPLC-MS/MS analysis of the isolated genomic DNA showed a clear dose-dependent integration of **10** into the genome, up to 1×10^{-3} per dN. We next searched for other 2'-fluorinated pyrimidine nucleosides and detected a dose-dependent presence of (*R*)-2'-F-dU **13** and (*R*)-2'-F-dT **14**, formed by deamination of **10** to **13** followed by methylation to **14** (Fig. 3c). In addition, we detected a dose-dependent formation of (*R*)-2'-F-mdC and (*R*)-2'-F-hmdC, confirming that compound **10** is biochemically converted as expected (Fig. 3c).

In order to quantify the levels of methylation, we synthesized the isotope labelled compounds [D$_3$]-F-mdC **18**, [^{15}N$_2$]-F-dC **19** and [^{15}N$_2$]-F-fdC **20** and used them as internal standards for quantification (Fig. 3a). Upon feeding mESCs with 1 µM **10** for three days, we detected around 3% (±0.5%) of **11** relative to **10**, which is similar to what is observed for the natural bases (Supplementary Fig. 5).
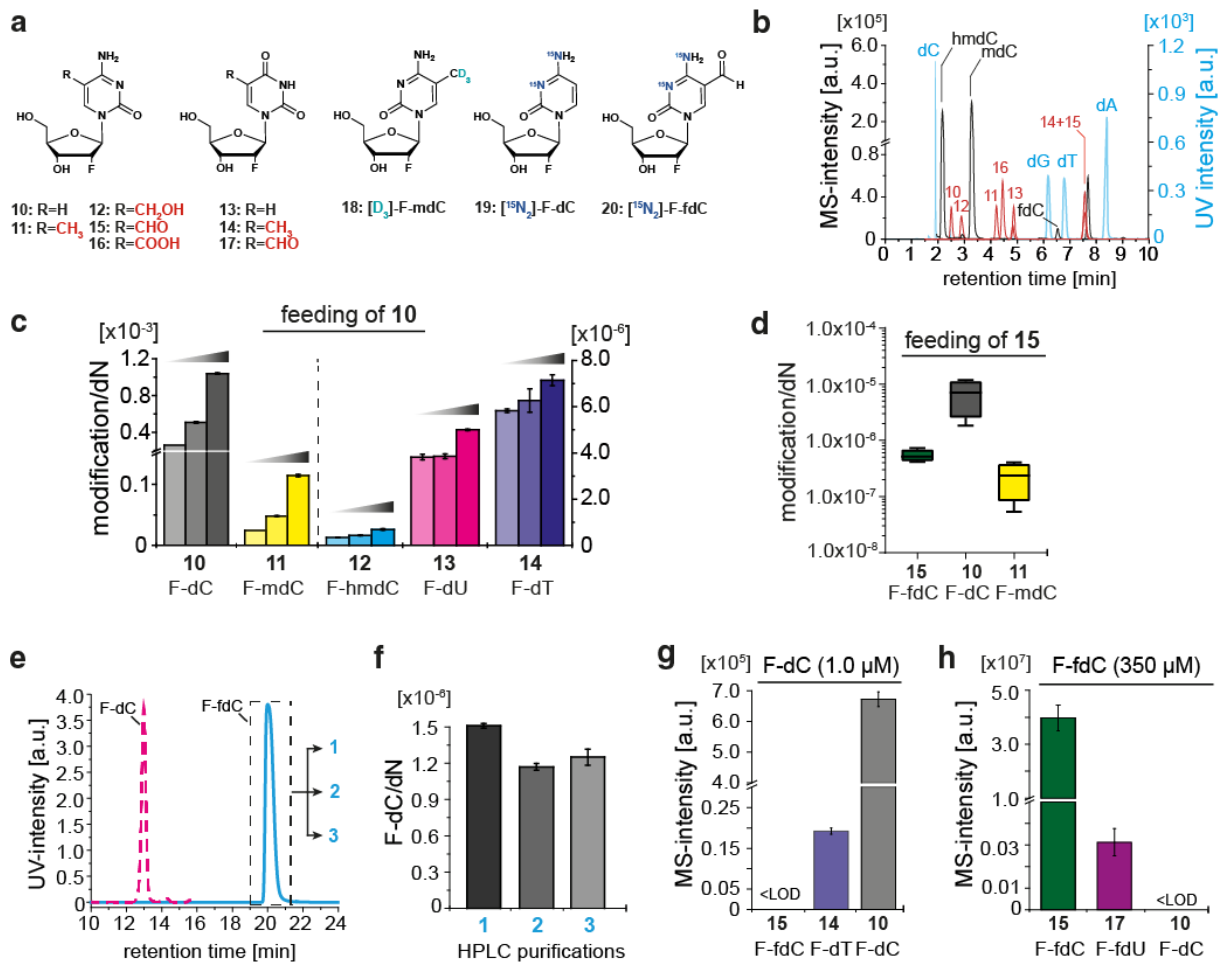
Figure 3 | F-fdC is converted into F-dC within the genome (a) Chemical structures of 2'-fluorinated dC and dU derivatives that were investigated and internal standards used. **(b)** Resulting UV- and MS-traces of the nucleosides under investigation. Light blue = canonical bases, black = natural dC derivatives, red = fluorinated bases. **(c)** Quantitative data of fluorinated pyrimidine derivatives after feeding **10** for three days at different concentrations (0.5 µM, 1.0 µM, 2.5 µM). A DNA sample from a single culture was measured as technical triplicates. **(d)** Quantitative data of fluorinated pyrimidine derivatives after feeding **15** at 350 µM for three days**.** Technical triplicates from four independent cultures were measured. **(e)** HPL-chromatogram of three consecutive purifications of **15** (1-3, blue line). The purple dashed line marks the position where a peak of contaminating F-dC would be expected. **(f)** Quantitative data after feeding of **15** that has been purified three consecutive times (from e). The levels of the deformylation product **10** remain the same ruling out any contribution from a possible contamination. **(g-h)** Quantitative data of fluorinated pyrimidine derivatives in the soluble pool after feeding **10** at 1 µM (g) and **15** at 350 µM (h) for three days. Technical triplicates from single cultures were measured. Technical triplicates from single cultures were measured. In (c, e-h) mean values with s.d. are shown.

(*R*)-2'-F-fdC is converted into (*R*)-2'-F-dC within the genome of mESCs

To study the direct C-C bond cleavage process, we again cultured mESCs in the presence of (*R*)-2'-F-fdC **15** (350 µM, 3 d), isolated the DNA and analysed the nucleoside composition. Next to genomic (*R*)-2'-F-fdC (5.7×10^{-7}/dN), we detected (*R*)-2'-F-dC at a level of 7.3×10^{-6}/dN (Fig. 3d). Because the nucleosides **13/14** were not detected, we suspected again that the observed reaction **15 → 10** occurs

8

directly within the genome. The detection limit of **13** and **14** is, however, around 5 fmol and so the compounds may just escape observation.

In order to substantiate the conclusion that genomic **15** undergoes intra-genomic C-C bond cleavage to **10**, we asked first if **15** could spontaneously deformylate. To investigate this possibility, an aqueous solution of **15** was heated to 60 °C for three days, but compound **10** was not detected. We also incubated **15** in culture medium for three days and were unable to detect any **10**. Finally, we added a 28mer oligonucleotide containing a single **15** to culture medium for three days, re-isolated the DNA-strand and searched for **10**. Formation of **10** was again not detected. Together these experiments exclude background deformylation.

We next analysed the possibility that **15** is contaminated with traces of **10**. To this end, the purity of **15** was checked by MS and indeed **10** was not found. In order to exclude the presence of even traces of **10** below the detection limit, we performed three consecutive HPLC purifications of **15**. This compound **15** is well separable from **10** because of a large retention time difference of 7.50 min, using our gradient (Fig. 3e). Feeding of the material **15** obtained from three consecutive purifications resulted in unchanged values of genomic **10**, arguing against the possibility that the detected **10** is an accumulated impurity (Fig. 3f).

To further substantiate that the C-C bond cleavage does not occur in the soluble nucleoside/nucleotide pool, we added **10** to the mESC culture for three days. UHPLC-MS/MS analysis of the soluble pool allowed us to detect **10**, **14** and in traces **13** (Fig. 3g). However, when we repeated the study with **15** (Fig. 3h), we detected just **15** in the soluble pool, plus the deaminated compound (*R*)-2'-F-dC **17**, but not **10**. We next determined the medium concentration of **10** that would be needed to reach the measured value for genome integrated **10** (7.3×10^{-6}/dN, Fig. 3d) and found that a concentration of 5-10 nM would be required (Supplementary Fig. 6). With a detection limit of 30 amol for **10** (40 μL injection), this is a concentration at which **10** is unambiguously detectable.

All these control experiments support the idea that the C-C bond cleavage to F-dC does occur within the genome. Interestingly, upon feeding of **15** we also detected the methylated derivative **11**, demonstrating that the demodified product **10** is methylated to **11** in the genome. Using the isotopically-labelled internal standards [D_3]-F-mdC, [^{15}N_2]-F-dC and [^{15}N_2]-F-fdC (Fig. 3a), we see that re-methylation of **10** provides levels of 2.8% (±0.3%) F-mdC (Fig. 3d), which is only slightly lower compared to what was observed when we fed **10** directly (Supplementary Fig. 5).

To study the time dependence of the C-C bond cleavage process, we fed compound **15** and measured the genome-integrated levels of **15**, **10** and **11**. Already at 0.5 h, we detected a stable incorporation of **15** (Fig. 4a). The C-C bond cleaved product **10** appears after about 1 h and here the levels increase

steadily (Fig. 4b). After about 4 h, we see the first re-methylated product **11** (Fig. 4c). If **10** were a contamination in the preparation of **15**, we would expect faster incorporation kinetics. When we fed both **15** and **10** simultaneously**,** a steady increase in the level of **10** was already observed after 5 min (Supplementary Fig. 7), confirming that our probe nucleosides are quickly incorporated into the genome. These data show that the C-C bond cleavage is a rapid process.
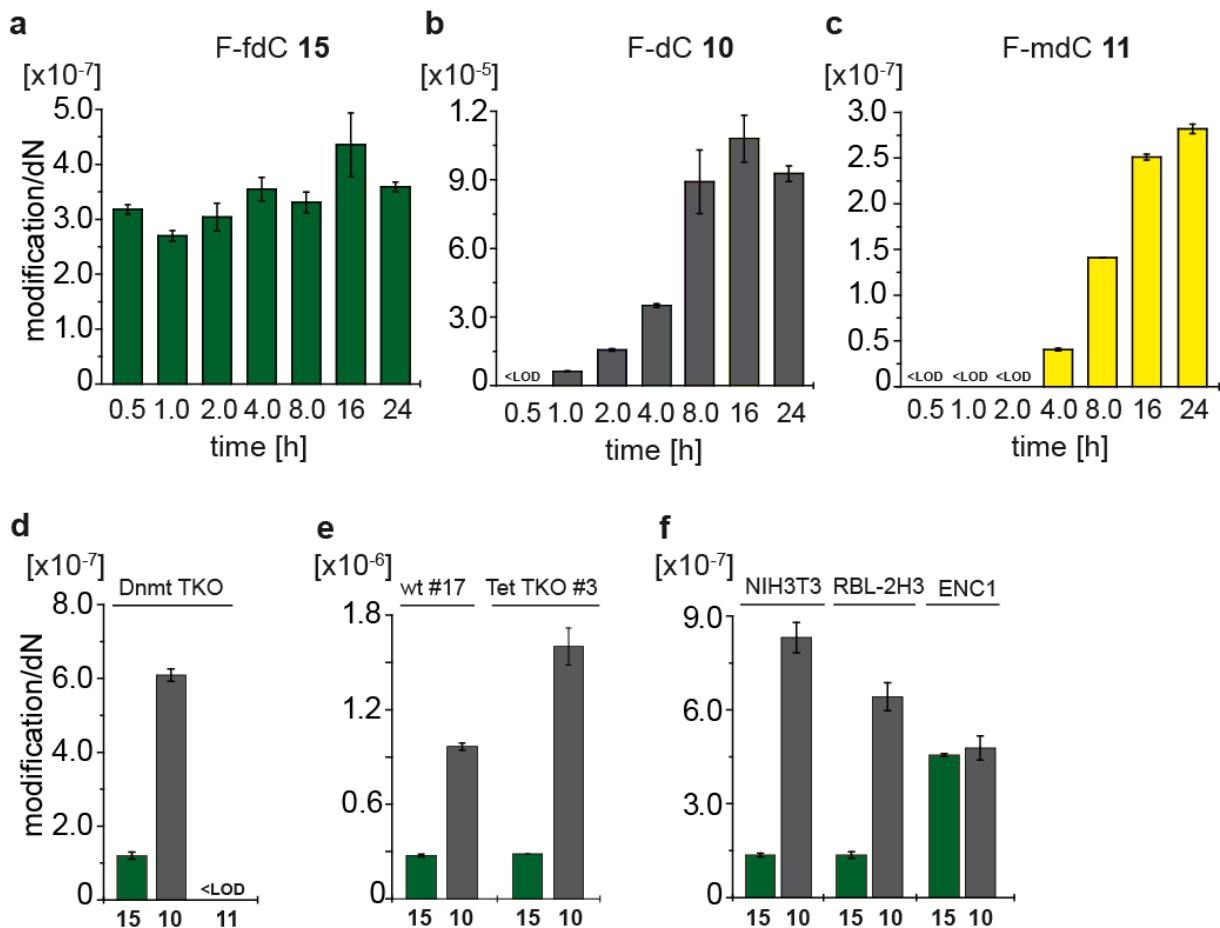

Figure 4 | Demodification of 2'-fluorinated fdC is a rapid process, does not require Dnmt or Tet enzymes and occurs also in somatic cell types. **(a-c)** Time course study showing the genomic build-up of F-fdC **15** (a), F-dC **10** (b) and F-mdC **11** (c) upon metabolic labelling with **15**. **(d-f)** Genomic levels of **15** and **10** upon metabolic labelling of Dnmt TKO mESCs (d), Tet TKO and corresponding wt mESC lines (e) and various somatic cell lines (f) with **15**. Genomic levels of **11** are also shown in (d). In all panels, mean values and s.d. of technical triplicate measurements from single representative experiments are shown. Data from two additional independent experiments are shown in Supplementary Figures 8 and 9.

Demodification of *(R)*-2'-F-fdC does not require Dnmt or Tet enzymes and occurs also in somatic cell types

In order to investigate if the re-methylation of **10** is driven by the known DNA methyltransferases and whether these are involved in demodification, we added **15** (350 µM, 3 d) to mESCs deficient of all active DNA methyltransferases Dnmt1, 3a and 3b (Dnmt triple knockout or TKO) and analysed the DNA.

In this experiment, the demodified product **10** is again detected, but the methylated product **11** is not seen (Fig. 4d and Supplementary Fig. 8), showing that DNA methyltransferases are responsible for methylation of **10** and are not required for demodification of **15**. We finally investigated if the Tet-enzymes are involved in demodification of **15**. Repeating the feeding experiment with mESCs lacking all three members of the Tet family (Tet triple knockout or TKO) showed however full demodification activity (Fig. 4e and Supplementary Fig. 9). The fact that the conversion **15 → 10** does not change in the absence of Tet proteins is particularly noteworthy. Because Tet enzymes were shown to accept **11** as a substrate converting it into **12, 15** and also (*R*)-2'-F-cadC **16**,[29] we can exclude that the observed C-C bond cleavage is in fact a Tet-dependent decarboxylation of **16**. Indeed, this result implies either that **15** is directly deformylated to **10** or that factors other than Tet enzymes can oxidise **15** to **16**, which is then decarboxylated to **10**. Ascorbic acid has been shown to increase Tet enzymatic activity *in vitro* and the levels of oxidised mdC derivatives *in vivo*.[30-32] Ascorbic acid treatment of mESC cultures fed with **15** indeed resulted in increased levels of naturally occurring fdC and cadC, but had no effect on conversion of **15** into unmodified product **10** (Supplementary Fig. 10). This further supports that demodification of **15** to **10** does not depend on the enzymatic activity of the Tet enzymes.

Finally, we tested whether conversion of **15** into **10** occurs in non-pluripotent cells by feeding **15** to cell lines representing a variety of cell types (Fig. 4f and Supplementary Fig. 11). Albeit to various degrees, we detected the **15 → 10** conversion in all these cell lines, arguing that it is rather widespread in mammalian cell types. In summary, our data prove that fdC is converted into dC *in vivo* through C-C bond cleavage and strongly suggest that this conversion is an intra-genomic process.

Discussion

In recent years, several mechanisms for active erasure of cytosine methylation from the genome have been proposed. Among these, the best-established mechanism entails Tet-mediated iterative oxidation of mdC to fdC or cadC, followed by the replacement of these higher oxidised derivatives with unmodified dC through BER. Considering the frequent occurrence of mdC in high-density clusters and its prevalent symmetrical configuration at CpG sites in vertebrate genomes, a BER-based erasure mechanism poses a substantial risk of creating clustered single and double strand breaks with potentially deleterious consequences. It is possible that excision of fdC and cadC by Tdg, processing of the abasic site and insertion of unmodified dC are orchestrated by a single multi-molecular complex, thus allowing tight control of strand breaks. Alternatively, it is also conceivable that, in order to minimize the potentially deleterious consequences of BER, complementary mechanisms are in place to remove fdC and cadC without involving DNA repair. In this context, it should be kept in mind that Tet3-dependent active demethylation of maternal and paternal genomes in the mouse zygote may not

require Tdg.[21] With an isotope-tracing experiment, using labelled dC-derived nucleosides in combination with highly sensitive UHPLC-MS/MS detection, we show here that in mammalian cells fdC is converted to dC while keeping the glycosidic bond intact. Evidence is provided that the C-C bond cleavage reaction happens when fdC is located inside the genome. This establishes an intra-genomic demodification process, independent of DNA repair. We also show that this process does not require any of the Tet-family enzymes. Therefore, unless other factors are able to oxidise fdC to cadC, this demodification process is likely a direct deformylation of fdC. Although we have firmly established the occurrence of a C-C bond cleavage of fdC/cadC to dC, the mechanism of this process remains to be defined, including the identification of the factors that mediate the demodification reactions.

We would like to emphasize that in our approach the probe nucleosides are randomly incorporated into the genome through DNA replication. Consequently, we cannot determine the sequence and genomic context where the demodification of fdC/cadC to dC takes place. Obviously, this would require a sequencing approach that allows the identification of the converted dC bases. In addition, as we could detect conversion of 2'-fluorinated fdC to 2'-fluorinated dC in mESCs and different somatic cell types. This indicates that the ability to carry out the demodification reaction may be widespread in mammalian cells and tissues, rather than being restricted to events of active genomic demethylation known to occur in specific developmental and tissue contexts. Assuming that the deformylation of fdC to dC establishes an active demethylation pathway, we need to emphasize that deformylation reactions and deformylases are widespread in nature. A prominent example is the enzyme Lanosterin-demethylase (CYP51A1), which oxidises the C-14α-methyl group of lanosterin to a formyl group to achieve deformylation under concomitant introduction of a double bond (dehydrating deformylation). This enzyme, a P450-type monooxygenase, contains a heme-cofactor that seems to utilise a nucleophilic Fe-peroxylanion species for attacking the substrate.[33,34] Another well studied enzyme that catalyses deformylation is the Aldehyde-Deformylating-Deoxygenase, which again uses a nucleophilic metal bond peroxylanion radical as the attacking species. This enzyme shortens fatty acid chains by oxidising the terminal methyl group to a formyl group, followed by deformylation.[35,36] In contrast to fdC, these deformylation reactions take place on formyl groups attached to saturated C-atoms, while in fdC, the formyl group is linked to an aromatic heterocycle. Such structures are known for decarboxylations and they are catalysed by the enzymes orotate[37] and isoorotate[38] decarboxylase. Indeed, it was suggested that the isoorotate decarboxylase could be a blueprint for a putatively existing cadC decarboxylase.[39] Our data now support the idea that fdC and possibly also cadC are converted to dC by a direct C-C bond cleavage. The questions as to when and where these reactions occur *in vivo* now requires the identification of putative catalytic factors.

Acknowledgements

Author contributions

Competing financial interest

The authors declare no competing financial interest.

References (main)

1   Tahiliani, M. *et al.* Conversion of 5-methylcytosine to 5-hydroxymethylcytosine in mammalian DNA by MLL partner TET1. *Science* **324**, 930-935, (2009).

2   Kriaucionis, S. & Heintz, N. The nuclear DNA base 5-hydroxymethylcytosine is present in Purkinje neurons and the brain. *Science* **324**, 929-930, (2009).

3   Pfaffeneder, T. *et al.* The Discovery of 5-Formylcytosine in Embryonic Stem Cell DNA. *Angew. Chem. Int. Ed.* **50**, 7008-7012, (2011).

4   Ito, S. *et al.* Tet proteins can convert 5-methylcytosine to 5-formylcytosine and 5-carboxylcytosine. *Science* **333**, 1300-1303, (2011).

5   He, Y. F. *et al.* Tet-mediated formation of 5-carboxylcytosine and its excision by TDG in mammalian DNA. *Science* **333**, 1303-1307, (2011).

6   Globisch, D. *et al.* Tissue distribution of 5-hydroxymethylcytosine and search for active demethylation intermediates. *PLoS One* **5**, e15367, (2010).

7   Münzel, M., Globisch, D. & Carell, T. 5-Hydroxymethylcytosine, the sixth base of the genome. *Angew. Chem. Int. Ed.* **50**, 6460-6468, (2011).

8   Pfaffeneder, T. *et al.* Tet oxidizes thymine to 5-hydroxymethyluracil in mouse embryonic stem cell DNA. *Nat. Chem. Biol.* **10**, 574-581, (2014).

9   Wagner, M. *et al.* Age-dependent levels of 5-methyl-, 5-hydroxymethyl-, and 5-formylcytosine in human and mouse brain tissues. *Angew Chem Int Ed Engl* **54**, 12511-12514, (2015).

10   Branco, M. R., Ficz, G. & Reik, W. Uncovering the role of 5-hydroxymethylcytosine in the epigenome. *Nat. Rev. Genet.* **13**, 7-13, (2011).

11   Wu, H. & Zhang, Y. Mechanisms and functions of Tet protein-mediated 5-methylcytosine oxidation. *Genes Dev.* **25**, 2436-2452, (2011).

12   Bachman, M. *et al.* 5-Formylcytosine can be a stable DNA modification in mammals. *Nat. Chem. Biol.* **11**, 555-557, (2015).

13   Su, M. *et al.* 5-Formylcytosine Could Be a Semipermanent Base in Specific Genome Sites. *Angew. Chem. Int. Ed.*, (2016).

14   Raiber, E. A. *et al.* 5-Formylcytosine alters the structure of the DNA double helix. *Nat. Struct. Mol. Biol.* **22**, 44-49, (2015).

15   Song, C. X. *et al.* Genome-wide profiling of 5-formylcytosine reveals its roles in epigenetic priming. *Cell* **153**, 678-691, (2013).

16   Kellinger, M. W. *et al.* 5-formylcytosine and 5-carboxylcytosine reduce the rate and substrate specificity of RNA polymerase II transcription. *Nat. Struct. Mol. Biol.* **19**, 831-833, (2012).

17   Zhu, C. *et al.* Single-Cell 5-Formylcytosine Landscapes of Mammalian Early Embryos and ESCs at Single-Base Resolution. *Cell Stem Cell*, (2017).

18   Hill, P. W., Amouroux, R. & Hajkova, P. DNA demethylation, Tet proteins and 5-hydroxymethylcytosine in epigenetic reprogramming: an emerging complex story. *Genomics* **104**, 324-333, (2014).

19   Wu, X., Inoue, A., Suzuki, T. & Zhang, Y. Simultaneous mapping of active DNA demethylation and sister chromatid exchange in single cells. *Genes Dev.* **31**, 511-523, (2017).

20   Maiti, A. & Drohat, A. C. Thymine DNA glycosylase can rapidly excise 5-formylcytosine and 5-carboxylcytosine: potential implications for active demethylation of CpG sites. *J. Biol. Chem.* **286**, 35334-35338, (2011).

21   Guo, F. *et al.* Active and Passive Demethylation of Male and Female Pronuclear DNA in the Mammalian Zygote. *Cell Stem Cell* **15**, 447-458.

22   Wu, S. C. & Zhang, Y. Active DNA demethylation: many roads lead to Rome. *Nat. Rev. Mol. Cell. Biol.* **11**, 607-620, (2010).

23   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**, 14593-14599, (2013).

24   Liutkeviciute, Z. *et al.* Direct decarboxylation of 5-carboxylcytosine by DNA C5-methyltransferases. *J. Am. Chem. Soc.* **136**, 5884-5887, (2014).

25   Schiesser, S. *et al.* Mechanism and stem-cell activity of 5-carboxycytosine decarboxylation determined by isotope tracing. *Angew. Chem. Int. Ed.* **51**, 6516-6520, (2012).

26   Jekunen, A. & Vilpo, J. A. 5-Methyl-2'-deoxycytidine. Metabolism and effects on cell lethality studied with human leukemic cells in vitro. *Mol. Pharmacol.* **25**, 431-435, (1984).

27   Vilpo, J. A. & Vilpo, L. M. Biochemical mechanisms by which reutilization of DNA 5-methylcytosine is prevented in human cells. *Mutat. Res.* **256**, 29-35, (1991).

28   Schröder, A. S. *et al.* Synthesis of (R)-Configured 2'-Fluorinated mC, hmC, fC, and caC Phosphoramidites and Oligonucleotides. *Org. Lett.* **18**, 4368-4371, (2016).

29   Schröder, A. S. *et al.* 2'-(*R*)-fluorinated mC, hmC, fC and caC triphosphates are excellent substrates for DNA polymerases and TET-enzymes. *Chem. Commun.* **52**, 14361-14364, (2016).

30   Blaschke, K. *et al.* Vitamin C induces Tet-dependent DNA demethylation and a blastocyst-like state in ES cells. *Nature* **500**, 222-226, (2013).

31   Minor, E. A., Court, B. L., Young, J. I. & Wang, G. Ascorbate induces ten-eleven translocation (Tet) methylcytosine dioxygenase-mediated generation of 5-hydroxymethylcytosine. *J. Biol. Chem.* **288**, 13669-13674, (2013).

32   Yin, R. *et al.* Ascorbic acid enhances Tet-mediated 5-methylcytosine oxidation and promotes DNA demethylation in mammals. *J. Am. Chem. Soc.* **135**, 10396-10403, (2013).

33   Hargrove, T. Y. *et al.* Substrate preferences and catalytic parameters determined by structural characteristics of sterol 14alpha-demethylase (CYP51) from Leishmania infantum. *J. Biol. Chem.* **286**, 26838-26848, (2011).

34   Lepesheva, G. I. *et al.* CYP51: A major drug target in the cytochrome P450 superfamily. *Lipids* **43**, 1117-1125, (2008).

35   Aukema, K. G. *et al.* Cyanobacterial aldehyde deformylase oxygenation of aldehydes yields n-1 aldehydes and alcohols in addition to alkanes. *ACS Catal.* **3**, 2228-2238, (2013).

36   Jia, C. *et al.* Structural insights into the catalytic mechanism of aldehyde-deformylating oxygenases. *Protein Cell* **6**, 55-67, (2015).

37   Fujihashi, M., Mnpotra, J. S., Mishra, R. K., Pai, E. F. & Kotra, L. P. Orotidine Monophosphate Decarboxylase--A Fascinating Workhorse Enzyme with Therapeutic Potential. *J. Genet. Genomics* **42**, 221-234, (2015).

38   Smiley, J. A., Angelot, J. M., Cannon, R. C., Marshall, E. M. & Asch, D. K. Radioactivity-based and spectrophotometric assays for isoorotate decarboxylase: identification of the thymidine salvage pathway in lower eukaryotes. *Anal. Biochem.* **266**, 85-92, (1999).

39   Xu, S. *et al.* Crystal structures of isoorotate decarboxylases reveal a novel catalytic mechanism of 5-carboxyl-uracil decarboxylation and shed light on the search for DNA decarboxylase. *Cell Res.* **23**, 1296-1309, (2013).

Online Methods

Chemical Synthesis

Synthetic schemes, detailed procedures and characterization of synthesized products can be found in the Supplementary Note. Unless noted otherwise, all reactions were performed using flame- or oven dried glassware under a atmosphere of nitrogen. Compounds **7** (*B.A.C.H. UG*) and **10** (*Carbosynth*) were commercially available. **15** was synthesized as previously described in the literature.[29] Identities of these compounds was confirmed by NMR and LC/MS-MS. Molsieve-dried solvents were used from *Sigma Aldrich* and chemicals were bought from *Sigma Aldrich*, *TCI*, *Carbolution* and *Carbosynth*. Technical grade solvents were distilled prior to extraction or chromatography of compounds. Reaction controls were performed using TLC-Plates from *Merck* (Merck 60 F_{254}), flash column chromatography purifications were performed on *Merck* Geduran Si 60 (40-63 µM). Visualization of the developed TLC plates was achieved through UV-absorption or through staining with *Hanessian's stain*. NMR spectra were recorded in deuterated solvents on *Varian VXR400S*, *Varian Inova 400*, *Bruker AMX 600*, *Bruker Ascend 400* and *Bruker Avance III HD*. HR-ESI-MS spectra were obtained from a *Thermo Finnigan* LTQ FT-ICR. IR-measurements were performed on a *Perkin Elmer Spectrum BX FT-IR* spectrometer with a diamond-ATR (Attenuated Total Reflection) unit. HPLC purifications were performed on a *Waters Breeze* system (2487 dual array detector, 1525 binary HPLC pump) using a Nucleosil VP 250/10 C18 column from *Macherey Nagel*, HPLC-grade MeCN was purchased from *VWR*.

Cell culture

Basal medium for mESC culture was DMEM high glucose containing 10% FBS, 2 mM L-glutamine, 100 U/mL penicillin, 100 µg/mL streptomycin, 1x MEM Non-essential Amino Acid Solution and 0.1 mM β-mercaptoethanol (all from *Sigma*). All mESC lines were maintained in naïve state on gelatin coated plates by supplementing basal medium with 1000 U/mL LIF (*ORF Genetics*), GSK3 inhibitor CHIR99021 at 3 µM and Mek inhibitor PD0325901 1 µM ("2i"). Metabolic labelling experiments with fluorine- or isotope-labelled nucleosides were performed by plating mESCs in priming conditions consisting of basal mESC medium supplemented with 3 µM CHIR99021 and Wnt pathway inhibitor IWR1-endo at 2.5 µM as previously reported[40]. Under these conditions primed cells remained pluripotent for at least seven days as determined by epifluorescence with an Oct4-YFP knock-in cell line[41]. Priming and labelling was performed for three days. Over this period naturally occurring genomic mdC and hmdC (Fig. 1b) reached levels similar to those recently reported for epiblast-like cells, which are regarded as the closest *in vitro* counterpart to non-committed post-implantation epiblast[42,43]. All inhibitors were purchased from *Selleckchem*. Dnmt TKO J1 mESCs were described in[44]. Two independent sets of Tet TKO and respective wt mESC lines were used: wt #17 and Tet TKO #3 were reported in[45] and wt #4 and Tet TKO #29 were described in[46]. J1 mESCs are from the 129/Sv/Jae strain, while all Tet TKO and corresponding wt mESC lines are from mixed genetic backgrounds.

The time course experiment was performed by culturing J1 mESCs under priming conditions for 48 h. The medium was exchanged to priming medium containing 350 µM F-fdC and cells were harvested after 0.5, 1.0, 2.0, 4.0, 8.0, 16 and 24 h, as described.

RBL-2H3, HeLa, NIH3T3 and Neuro-2a cells were cultured in DMEM high glucose containing 10% FBS, 2 mM L-glutamine, 100 U/ml penicillin, 100 µg/ml streptomycin. CHO-K1 cells were maintained in DMEM/F-12 supplemented as reported above for the other somatic cell lines. ENC1 neural stem cells were cultured as previously described[47]. Cells were exposed to labelled nucleosides for four (RBL-2H3 and NIH3T3), five (CHO-K1), six (Neuro-2a) and seven days (HeLa and ENC1).

Labelled nucleosides were added to the culture medium at the following concentrations: F-fdC (**15**), 350 µM; F-dC (**10**), 0.5, 1.0 and 2.5 µM; $[^{13}C_5][^{15}N_2]$-fdC (**1**), 50 µM; $[^{13}C_9][^{15}N_3]$-dC (**7**), purchased from *B.A.C.H. UG*), 100 µM.

Isolation of genomic DNA

Cultures were washed with PBS and lysed by adding RLT buffer (Qiagen) containing 400 µM each of 2,6-di-tert-butyl-4-methylphenol (BHT) and desferoxamine mesylate (DM) directly to the plates. Isolation of genomic DNA was performed with Zymo-Spin V, V-E or IIC-XL columns according to the instruction of the ZR-Duet DNA/RNA MiniPrep Kit (Zymo Research) with the following modifications. DNA was sheered by bead milling in 2 mL microfuge tubes using one 5 mm diameter stainless steel bead per tube and a MM400 bead mill (Retsch) set at 30 Hz for 1 min. Lysates were then loaded onto spin columns and the bound material was first incubated for 15 min with Genomic Lysis Buffer (Zymo Research) supplemented with 0.2 mg/mL RNase A (Qiagen). After washing genomic DNA fragments were eluted with water containing 0.4 µM of each BHT and DM.

DNA digestion

0.5–10 µg of genomic DNA in 35 µL H_2O were digested as follows: An aqueous solution (7.5 µL) of 480 µM $ZnSO_4$, containing 42 U nuclease S1 (*Aspergillus oryzae*, Sigma-Aldrich), 5 U Antarctic phosphatase (New England BioLabs) and specific amounts of labeled internal standards were added, and the mixture was incubated at 37 °C for 3 h. After addition of 7.5 µl of a 520 µM $[Na]_2$-EDTA solution, containing 0.2 U snake venom phosphodiesterase I (*Crotalus adamanteus*, USB corporation), the sample was incubated for 3 h at 37 °C or overnight and then stored at −20 °C. Prior to UHPLC-MS/MS analysis, samples were filtered by using an AcroPrep Advance 96 filter plate 0.2 µm Supor (Pall Life Sciences).

LC/MS-MS analysis of DNA samples.

Quantitative UHPLC-MS/MS analysis of digested DNA samples was performed using an Agilent 1290 UHPLC system equipped with a UV detector and an Agilent 6490 triple quadrupole mass spectrometer.

Prior to every measurement series, external calibration curves were measured in order to quantify the levels of the F-nucleosides (Supplementary Fig. 12). Additionally, [$^{15}N_2$]-F-dC (**19**), [$^{15}N_2$]-F-hmdC (**37**) and [$^{15}N_2$]-F-fdC (**20**) were used to validate the resulting peaks by co-injection. For exact quantification of fluorinated nucleosides also internal quantification with stable isotope dilution techniques for F-fdC, F-dC and F-mdC were developed (Supplementary Fig. 13). Natural nucleosides were quantified with the stable isotope dilution technique. An improved method, based on earlier published work[23,25,48-50] was developed, which allowed the concurrent analysis of all nucleosides in one single analytical run.[8] The source-dependent parameters were as follows: gas temperature 80 °C, gas flow 15 L/min (N$_2$), nebulizer 30 psi, sheath gas heater 275 °C, sheath gas flow 11 L/min (N$_2$), capillary voltage 2,500 V in the positive ion mode, capillary voltage −2,250 V in the negative ion mode and nozzle voltage 500 V. The fragmentor voltage was 380 V/ 250 V. Delta EMV was set to 500 (positive mode) and 800 (negative mode). Compound-dependent parameters are summarized in Supplementary Tables 1 - 4. Chromatography was performed by a Poroshell 120 SB-C8 column (Agilent, 2.7 µm, 2.1 mm × 150 mm) at 35 °C using a gradient of water and MeCN, each containing 0.0085% (v/v) formic acid, at a flow rate of 0.35 mL/min: 0 →4 min; 0 →3.5% (v/v) MeCN; 4 →7.9 min; 3.5 →5% MeCN; 7.9 →8.2 min; 5 →80% MeCN; 8.2 →11.5 min; 80% MeCN; 11.5 →12 min; 80 →0% MeCN; 12 →14 min; 0% MeCN. The effluent up to 1.5 min and after 12 min was diverted to waste by a Valco valve. The autosampler was cooled to 4 °C. The injection volume was amounted to 39 µL.

Quantification of nucleosides

Prior to every sample set, calibration curves to quantify all fluorine labelled nucleosides were measured under the same conditions and settings. All calibration curves are valid within the range of 1-500 fmol with five measuring points and measured as technical triplicates. Supplementary Figure 12 shows representative calibration curves for all Fluoro-nucleosides used for the quantification.

In order to obtain the internal calibration curves for exact quantification, each standard, namely [$^{15}N_2$]-F-fdC (**20**), [$^{15}N_2$]-F-dC (**19**) and [D_3]-F-mdC (**18**) was analysed in comparison to the corresponding non-labelled nucleoside with constant concentration. Technical triplicates were measured and the linear regression was applied using Origin® 6.0 (Microcal™). Therefore, the ratio of the area under the curve of unlabelled nucleoside (A) to the labelled standard (A*) was plotted against the ratio of the amount of unlabelled nucleoside (n) to the labelled one (n*) (see Supplementary Fig. 13). Acceptable precision (< 20% relative s.d. within each triplicate) and accuracy (80-120%) was achieved for all three calibration curves. The accuracy is calculated as the ratio of the measured to the calculated ratios of the areas (A/A*) under the curves in percent. The ratios of the areas (A/A*) can be calculated by using the linear equations for the corresponding ratio of amount (n/n*). The lower limit of detection was defined as the detected amount, which is three times higher than the blank response (LOD). The lower limit of detection (LLOQ) and the upper limit of detection (ULOQ) are the lowest, respectively the highest

amount (n) and ratio of the amounts (A/A*) fulfilling the requirements of the corresponding linear equation.

Nucleoside stability test

Compounds **1** and **15** were incubated at 100 µM in mESC culture medium at 37 °C and 5% CO_2 for 3 d. For the recovery of the nucleosides *Supel-Select SPE HLB* cartridges from Sigma Aldrich were used. Prior to use, these cartridges were equilibrated with MeOH, H_2O and diluted HCl (pH = 4). The pH of the samples was adjusted to 4 and the acidic solution was loaded on the cartridges. After washing with 10 mL of H_2O, the cartridges were dried *in vacuo.* The nucleosides were eluted with MeOH/MeCN (1:1), evaporated to dryness *via* speedvac and resuspended in H_2O.

Oligonucleotide stability test

An oligonucleotide (6.9 pmol) containing one F-fdC (28mer) was incubated in mESC culture medium at 37 °C and 5% CO_2 for 3 d. For the recovery of the oligonucleotide *Oligo Clean & Concentrator* from Zymo Research was used according to the manual. The resulting oligonucleotide was dissolved in H_2O and digested as described for genomic DNA.

Extraction of nucleoside/nucleotide soluble pools

J1 mESCs were plated under priming conditions (as described above) for three days. The culture medium was supplemented with 1.0 µM F-dC (**10**) or 50 µM [$^{13}C_5$][$^{15}N_2$]-fdC (**1**) or 350 µM F-fdC (**15**). Cells were washed twice with PBS (Sigma Aldrich), harvested by trypsinization and pelleted by centrifugation for 3 min at 300 g. 500 µL ice cold 50% (v/v) acetonitrile was added dropwise to the pellet and vortexed.[51] The mixture was incubated on ice for 10 min. The insoluble fraction was then separated from the soluble pool by centrifugation for 10 minutes at 21000 x g at 0 °C. The supernatant was removed and used for nucleoside isolation. The soluble fraction containing the nucleosides was dried by lyophilization and metabolites were purified using *Supel-Select SPE HLB* cartridges (as described in the Nucleoside stability test) prior to UHPLC-MS/MS analysis.

Data availability and code availability statements

The data that support the findings of this study are available from the corresponding authors upon reasonable request.

Methods-only References

40   Kim, H. *et al.* Modulation of beta-catenin function maintains mouse epiblast stem cell and human embryonic stem cell self-renewal. *Nat. Commun.* **4**, 2403, (2013).

41   Toyooka, Y., Shimosato, D., Murakami, K., Takahashi, K. & Niwa, H. Identification and characterization of subpopulations in undifferentiated ES cell culture. *Development* **135**, 909-918, (2008).

42   Shirane, K. *et al.* Global Landscape and Regulatory Principles of DNA Methylation Reprogramming for Germ Cell Specification by Mouse Pluripotent Stem Cells. *Dev. Cell* **39**, 87-103, (2016).

43   Hayashi, K., Ohta, H., Kurimoto, K., Aramaki, S. & Saitou, M. Reconstitution of the mouse germ cell specification pathway in culture by pluripotent stem cells. *Cell* **146**, 519-532, (2011).

44   Tsumura, A. *et al.* Maintenance of self-renewal ability of mouse embryonic stem cells in the absence of DNA methyltransferases Dnmt1, Dnmt3a and Dnmt3b. *Genes Cells* **11**, 805-814, (2006).

45   Hu, X. *et al.* Tet and TDG mediate DNA demethylation essential for mesenchymal-to-epithelial transition in somatic cell reprogramming. *Cell Stem Cell* **14**, 512-522, (2014).

46   Dawlaty, M. M. *et al.* Loss of Tet enzymes compromises proper differentiation of embryonic stem cells. *Dev. Cell* **29**, 102-111, (2014).

47   Liu, N. *et al.* Intrinsic and extrinsic connections of Tet3 dioxygenase with CXXC zinc finger modules. *PLoS One* **8**, e62755, (2013).

48   Cao, H. & Wang, Y. Collisionally activated dissociation of protonated 2'-deoxycytidine, 2'-deoxyuridine, and their oxidatively damaged derivatives. *J. Am. Soc. Mass Spectrom.* **17**, 1335-1341, (2006).

49   Spruijt, C. G. *et al.* Dynamic readers for 5-(hydroxy)methylcytosine and its oxidized derivatives. *Cell* **152**, 1146-1159, (2013).

50   Wang, J. *et al.* Quantification of oxidative DNA lesions in tissues of Long-Evans Cinnamon rats by capillary high-performance liquid chromatography-tandem mass spectrometry coupled with stable isotope-dilution method. *Anal. Chem.* **83**, 2201-2209, (2011).

51   Dietmair, S., Timmins, N. E., Gray, P. P., Nielsen, L. K. & Krömer, J. O. Towards quantitative metabolomics of mammalian cells: Development of a metabolite extraction protocol. *Anal. Biochem.* **404**, 155-164, (2010).