5-Formylcytosine could be a semi-permanent base in specific genome sites

Meng Su, Angie Kirchner, Samuele Stazzoni, Markus Müller, Mirko Wagner, Arne Schröder and Thomas Carell*[a]

Center for Integrated Protein Science at the Department of Chemistry. Ludwig-Maximilians-Universität München, Butenandtstrasse 5–13, 81377 München (Germany), E-mail: thomas.carell@lmu.de Homepage: http://www.carellgroup.de

Published 25.08.2016 in Angewandte Chemie International Edition, https://doi.org/10.1002/anie.201605994

Abstract: 5-Formyl-2'-deoxycytosine (fdC) is a recently discovered epigenetic base in the genome of stem cells, with yet unknown functions. Sequencing data show that the base is enriched in CpG islands of promoters and hence likely involved in the regulation of transcription during cellular differentiation. fdC is known to be recognized and excised by the enzyme thymine-DNA-glycosylase (Tdg). As such, fdC is believed to function as an intermediate during active demethylation. In order to understand the function of the new epigenetic base fdC, it is important to analyze its formation and removal at defined genomic sites. Here, we report a new method that combines sequence-specific chemical derivatization of fdC with droplet digital PCR that enables such analysis. We show initial data, indicating that the repair protein Tdg removes only 50% of the fdCs at a given genomic site, arguing that fdC is a semi-permanent base.

DNA contains besides the sequence information a second, epigenetic information level, which encodes how actively the controlled gene is transcribed.[1] Today, next to the four canonical bases, four additional epigenetic bases are known.[2] These are 5-methylcytosine (5mC), 5-hydroxymethylcytosine (5hmC),[3] 5-formylcytosine (5fC),[4] and 5-carboxycytosine (5caC).[5] (Figure 1a) Over the last years, sensitive mass spectrometry-based methods have helped to reveal the global levels of these epigenetic bases in stem cells[4,6] and tissues including the brain.[7] In order to learn about the levels and the distribution of the epigenetic bases at specific sites in the genome, different sequencing methods were developed[8] in which selective chemical derivatization of the bases is performed[9], sometimes in combination with bisulfite sequencing.[9c,10] Although these methods provide information about the distribution of the bases at a given time point, it is a hallmark of epigenetic information that it changes dynamically. To gain deeper insight into the dynamics of the epigenetic information layer at a single position in the genome, it is therefore essential to develop methods that allow following the changes of, for example, fdC at a specific location in the genome over time.[11] A perfect method will ultimately allow parallel monitoring of fdC dynamics at different genomic sites.

The central question addressed in this manuscript is: Are the measured global data of the past averages from different processes at different positions in the genome, or do they reflect what is happening at an individual site in the genome. To answer this question, we developed a sequence specific chemical derivatization method that allows in combination with droplet digital PCR to monitor the epigenetic base fdC at different loci directly in the genome of stem cells.

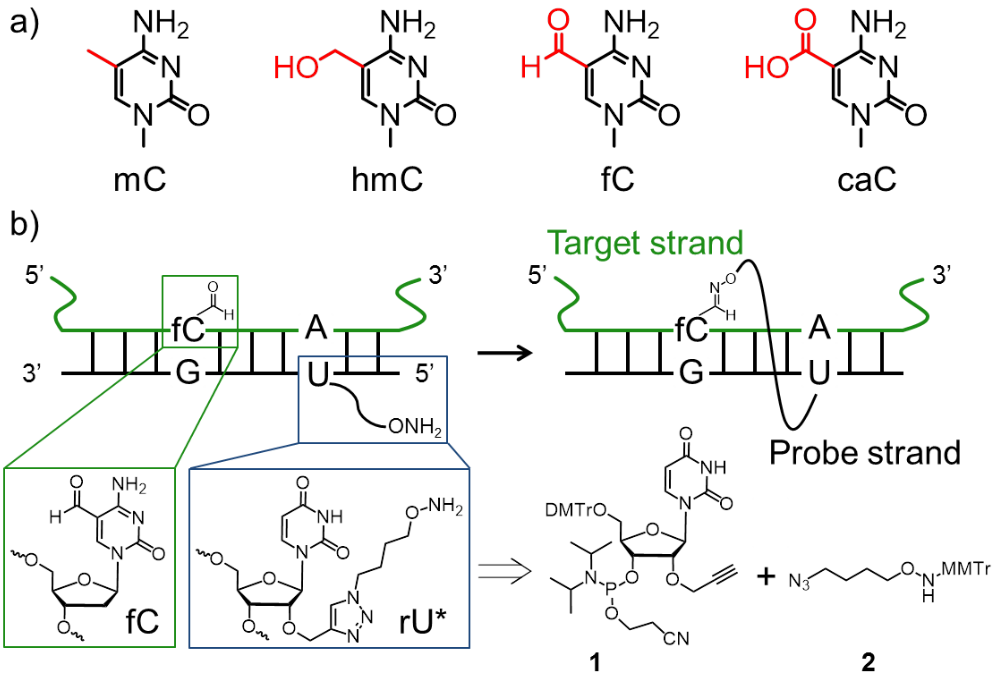


Figure 1. a) Structures of cytosine epigenetic modifications; b) Schematic representation of the fdC detection strategy, and used building blocks including the click chemistry-based assembly of the rU* probe molecule.

For the sequence specific localization of fdC in the genome, we utilize a small probe oligonucleotide (Figure 1b, Table S1), which contains a hydroxylamine tether that is able to form a covalent linkage with fdC so that the probe strand is subsequently tightly bound to the target.[4] We examined systematically different linker lengths, linker attachment points and distances. Best results were obtained when we incorporated the 2'-*O*-propargyl uridine using its phosphoramidite **1** into the probe oligonucleotide and attached the azido-C4-hydroxylamine **2** using the Cu(I)-catalyzed version (click reaction) of the Huisgen-reaction.[12] We protected the hydroxylamine unit for the click reaction with a monomethoxytrityl group (MMTr), which was cleaved afterwards with acetic acid at 25°C. This brief exposure of the probe oligonucleotide to acidic conditions did not cause significant depurination. After solid-phase synthesis, click modification of the oligonucleotide and a final purification step (Figure S1), we obtained oligonucleotides with different sequences and lengths containing an rU-hydroxylamine base (rU*) at different positions for reaction with the fdC-base on the target strand. For the following experiments, we prepared 13-mer long oligonucleotides.

To investigate at which position the linker in the probe strand would react best with fdC in the target strand, we varied the position of rU* relative to fdC and explored different reaction conditions (data not shown). Excellent results were finally obtained when probe strand **P1,** containing rU* exactly 4 basepairs in 5' direction relative to fdC, was hybridized to the fdC target strand **T1** in the presence of catalytic amounts of 4-methoxyaniline (Fig 2a). With this catalyst, the crosslinking reaction is complete after 24 h with yields exceeding 95%. Without the catalyst, only about 50% yield could be obtained (Fig 2b).

In order to increase the rate of the reaction, we tested other catalysts. We observed the best results when we used 1,4-diaminobenzene as a catalyst, in which case the crosslinking reaction between **T1** and **P1** is completed already after 3 h (Figure 2c). Duplex formation (**T1**:**P1**) was analyzed using denaturing PAGE and quantified by fluorescence (Figure 2d).

When fdC is located one base pair further away from rU* without changing the probe strand, we observe slower reaction (Figure S2). These results show that rU* placed four or five bases away from fdC in 5'-direction to fdC allows the tether to reach the formyl group of fdC via the major groove of the duplex (Figure S3).

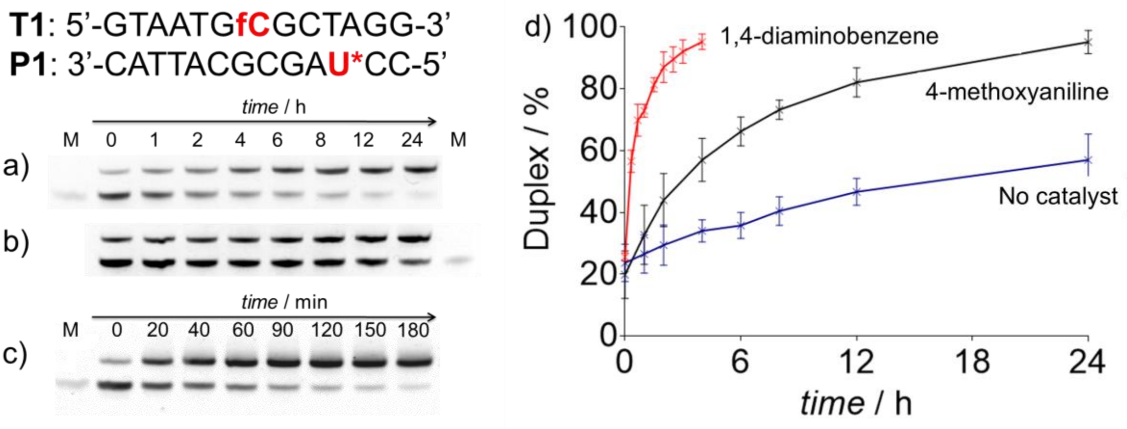


Figure 2. Denaturing PAGE gel showing the duplex formation between T1 and P1 at 25°C: a) with the catalyst 4-methoxyaniline; b) without a catalyst; c) with the catalyst 1,4-diaminobenzene; d) Quantification of the DNA duplex formation during the reaction. Black: catalyst 4-methoxyaniline, blue: no catalyst, red: catalyst 1,4-diaminobenzene. Error bars represent the standard error of the mean calculated from three replicates. Conditions: 2 µM oligonucleotides, 100 mM NaCl, 10 mM NaOAc buffer pH 6.0, 10 mM 4-methoxyaniline. M = single strand marker. The time point 0 is after re-annealing.

MALDI-TOF data confirmed that the crosslinks form as expected (Figure S4). For the reacted duplex **T1**:**P1**, we obtained the correct molecular weight for the duplex with m/z_{found} = 8081.9 (m/z_{calc} = 8084.7). As expected, the oxime formation reaction between **T1** and **P1** leads to a higher melting temperature of the hybridized and reacted duplex (Figure S5). Typically, we observed that the un-crosslinked 13-mer duplex melts at around 44°C. The duplex after crosslink formation shows a melting temperature of above 80°C.

Because pyrimidine bases are able to react with nucleophiles also at the C6 position in a Michael-type reaction, which is the basis for bisulfite sequencing, we next tested if the reaction of rU* is possible with other pyrimidines (Figure 3). To our delight, hybridization of the rU*-containing probe strand with target strands containing dT, dC, mdC, hmdC and cadC (**T3-8**) gave no reaction. Reaction is, however, observed with abasic sites. This is important because fdC and cadC are substrates for base excision repair and hence could in principle be precursors for abasic sites.[13] In this sense, rU* always reports the presence of fdC and also potentially of fdC and cadC derived abasic sites.

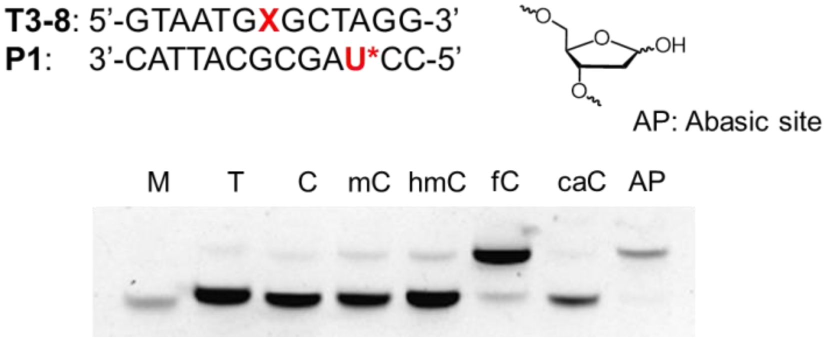


Figure 3. Denaturing PAGE gel showing duplex formation of T3-8 and P1 at 25°C after 24 h.

We finally turned the sequence specific fdC detection possibility into a method for detecting single fdC bases at a defined position in whole genomes. To this end, we coupled the chemistry to droplet digital PCR[14]-based amplification and readout.

Genomic target DNA (**Tg**) was in the first step isolated from mouse embryonic stem cells (mESCs) at different time points during priming from naïve cells. We also isolated genomic DNA from mESCs with a knockout of the Tdg repair enzyme (Tdg$^{-/-}$) to block excision and repair of fdC and cadC. We finally also isolated genomic DNA from mESCs lacking any of the three methyltransferases (Dnmt1, 3a and 3b). These stem cells lack mdC and are hence unable to produce the oxidized xdC (x = hm, f and ca) epigenetic bases. This genomic DNA served in our studies consequently as a negative control. For analysis, we selected two different fdC sites that were reported to have high fdC contents.[10c] We focused initially on the 30,020,539th site of chromosome 16 *Mus musculus* (MM9) located on the exon 3 of 632428C04Rik. It was found to contain 23% of fdC based on redBS-Sequencing. The second site we studied was the 8,846,677th site of chromosome 15 which is located in non-coding DNA. This site was reported to contain 32% of fdC.

For the first site, we reacted a 25-mer probe (**P2**, SI) containing the rU* base with **Tg** using 1,4-diaminobenzene as catalyst. In the absence of fdC, a covalent bond between **P2** and **Tg** cannot form. To remove the excess of probe, we loaded the **Tg:P2** complex onto an NEB Monarch DNA cleanup column and rinsed the column with wash buffer to elute oligonucleotides shorter than 50-mer, which is the unbound **P2**. After this washing, we eluted the **Tg:P2** with TE buffer. UV/Vis analysis of the eluted material showed a typical gDNA spectrum. We next added a 70-mer 5'-phosphorylated reporter strand (**R1**, SI) which hybridizes with an 18-nt stretch directly adjacent to the probe strand and ligated both probe and the reporter at 60°C by addition of Ampligase to form **R1-P2** as depicted schematically in Figure 4a.

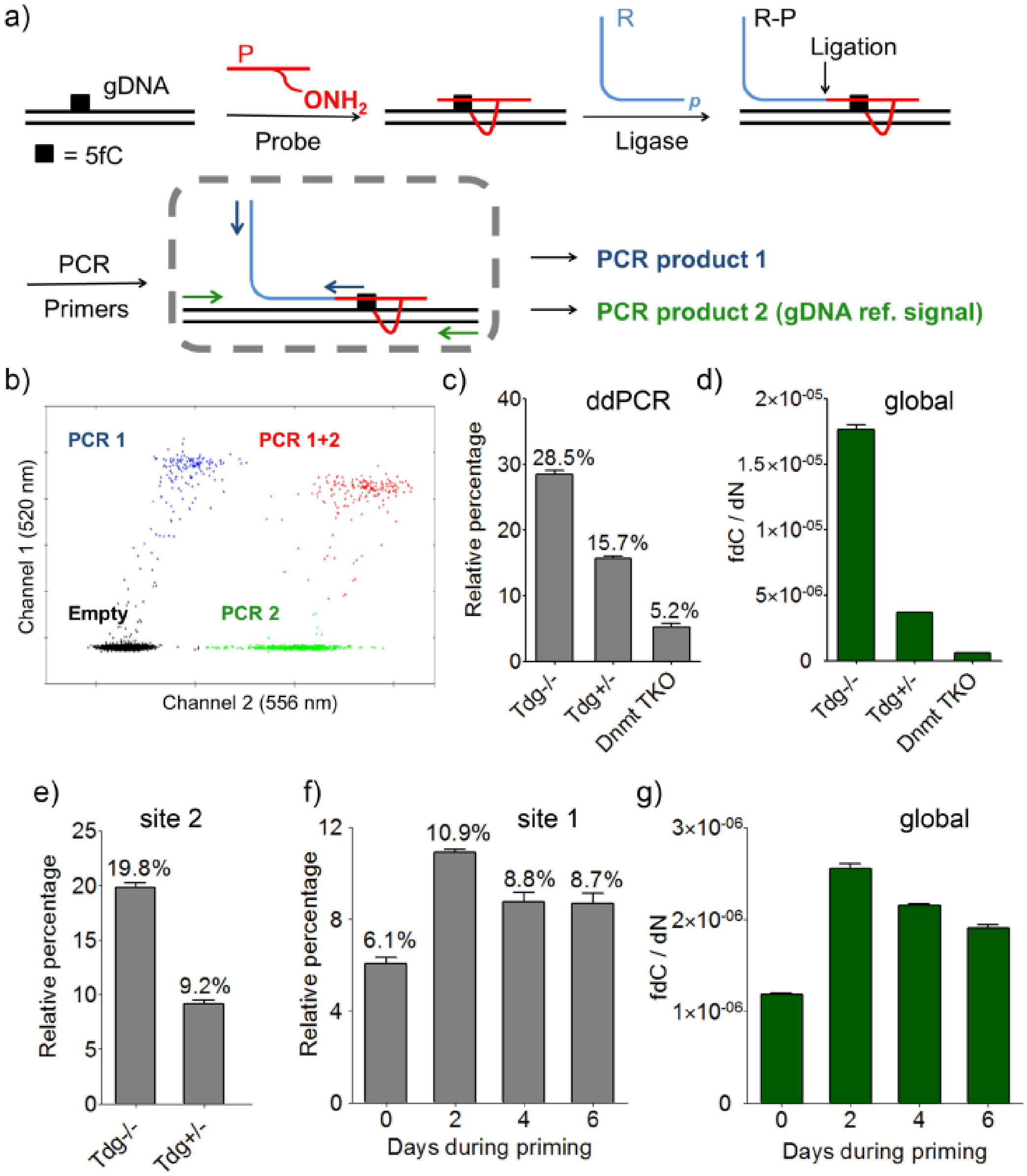


Figure 4. The fdC sequencing method: a) Schematic representation of the method, black line: gDNA; red segments: fdC probe; blue segments: reporter strands; arrows: PCR primer pairs; b) Typical 2-D plot of droplet fluorescence; c) Cluster ratios for position 1 in Tdg-/-, Tdg+/- and Dnmt TKO mES cells; d) Global fdC quantification in Tdg-/-, Tdg+/- and Dnmt TKO mES cells using our LC-MS method; e) Cluster ratios for position 2 in Tdg-/- and Tdg+/- mES cells; f) Cluster ratios for position 1 in wild-type mES cells at different days after priming; g) Global quantification data for the wild-type mES cells measured by LC-MS.

We next added two sets of primers to the assay (blue and green arrows, Figure 4a) to amplify the ligation product relative to the target duplex. Importantly, the blue primers recognize only the hybrid **R1-P2** probe generated in the ligation step while the green primers indicate the presence of gDNA. The amplification was monitored with two different TaqMan probes which showed fluorescence at 520 and 556 nm. This relative detection is needed to normalize on the amount of input gDNA. Because conventional real-time PCR is known to become inaccurate when copy number differences less than the 10-fold need to be resolved, we used droplet digital PCR. In this method, small droplets are generated with one droplet containing a maximum of one of the fully assembled analysis constructs shown in Figure 4a. The PCR reaction takes place in the droplets, producing a specific signal. Subsequent color-counting of each individual droplet yields numbers from which one can accurately calculate the amount of fdC, even if the fdC values are very low. A representative plot of the data is given in Figure 4b. Empty drops give no PCR signal (black dots in Figure 4b). Drops containing only **Tg** give only the PCR signal from the green primers (green dots in Figure 4b). Blue dots are obtained due to the dissociation of the ligated product **R1-P2** from **Tg** in the ligation process which is performed at 60°C for 10 h. The red signals are finally generated from droplets that contain both PCR products. For the calculation, please see the Supporting Information.

Using the method, we first studied mES cells lacking the Tdg enzyme (Tdg$^{-/-}$). A rather high level of 28.5% fdC was measured at the first locus (Figure 4c) in agreement with the results from redBS-seq.[10c] When we performed the study, however, with mESCs having an active Tdg repair enzyme (Tdg$^{+/-}$) we measured that the fdC level drops at this particular position to 15.7% (Figure 4c). This is very important because it shows that Tdg removes only half of the fdCs at a given site and also unusual due to the fact that repair glycosylases are known to find basically all possible substrates. The result underpins the high dynamics of fdC formation and repair at a given site. When we studied the fdC content at this location in mESCs lacking any methyltransferase (Dnmt TKO) the fdC level drops as expected to a little more than 5%, showing that the reported levels of fdC in the Tdg$^{+/-}$ cells are real and not an artifact. In order to elucidate if single-site fdC levels (Figure 4c) follow global genomic fdC levels, we quantified the total levels of fdC in these cells (Figure 4d). These global data are in good agreement with the data obtained from single site fdC quantifications. Thus, our new data make a scenario where fdC is fully removed at one site and shielded from repair at another place unlikely. Instead, fdC is even at a given position only partially removed in a cell population. Alternatively, it may be that Tdg removes fdC differently on the two chromosomes, which however needs further investigation.

In order to verify the data, we repeated the Tdg study at a second genomic site (8,846,677th nucleoside of chromosome 15). For this site, we designed a new probe strand **P3** and a new reporter strand **R2** and performed again ddPCR with two sets of primers (Figure S5). Comparing the data obtained from Tdg$^{-/-}$ cells with the data from Tdg$^{+/-}$ cells, again only a 50% reduction of the fdC level is shown at this position, in full agreement with the data above obtained from the first position (Figure 4e).

We finally performed a kinetic study in which we monitored the fdC development at the first position during priming of stem cells (Figure 4f). We see that the fdC levels rise at the given position with a strong increase in the early phase of priming, followed by a small decline phase and finally stable values (Figure 4f) again in agreement with the global data that we again measured using our reported method (Fig 4g).

The fact that our method is providing the same trends as seen in the global data at a single genomic site makes us confident that our method is robust and reliable reporting what happens at an individual site. Because single-site and global data go in parallel, we have now first evidence that the reported global trends are reflecting what happens at each individual fdC site, rather than evening out largely different dynamics at separate sites. Another interesting result of this study is that the repair enzyme Tdg removes only half of the fdC bases at a given genomic site in an mESC population, which argues that fdC is a semi-permanent base at a given position in the genome.

Experimental Section

Probe crosslinking gDNA solution (1.2 µg), fdC probe (1 µM, 2 µL), NaH_2PO_4-Na_2HPO_4 buffer (200 mM, pH = 6.0, 2 µL), NaCl aq. (1.5 M, 2µL), and ddH_2O were mixed to a final volume of 18 µL. The mixture was heated to 95°C for 3 min and then cooled down rapidly to 25°C. 1,4-Benzenediamine aq. (10 mM, 2 µL) was added and the reaction vial was shaken for 6 h at 25°C. The mixture was neutralized with Na_2HPO_4 aq. (200 mM, 40 µL) before purification with the NEB Monarch PCR DNA Cleanup Kit.

Ligation The above described gDNA solution (300 ng), reporter strand (20 nM, 1 µL), Ampligase reaction buffer (10×, 2 µL), Ampligase from Epicentre (5 U/µL, 2 µL, 10 U) and ddH_2O were mixed to a final volume of 20 µL. The mixture was heated to 95°C for 3 min, and then 94°C for 1 min, 60°C for 1 h and back to 94°C for 10 cycles. Then, the reaction mixture was diluted with Tris-HCl buffer (200 mM, pH = 7.6, 50 µL) before purification using the NEB Monarch PCR DNA Cleanup Kit.

Droplet digital PCR ddPCR was conducted on a Bio-Rad QX100 ddPCR System. For one reaction, gDNA (6 ng), four primers (18 µM each, 1 µL), two TaqMan probes (5 µM each, 1 µL), digital PCR Supermix for Probes (no dUTP, 2×, 10 µL), and ddH_2O were mixed to a final volume of 20 µL. PCR cycle: 95°C for 10 min, 94°C for 30 sec and 64°C for 1 min for 35 cycles, then 98°C for 10 min and cooled down to 12°C, with a temperature ramp of 2°C/s. For a detailed description please see the Supporting Information.

Acknowledgements

We thank K. Hufnagel for preparing the phosphoramidites of the epigenetic bases. This project has received funding from Deutsche Forschungsgemeinschaft SFB1032 and the Excellence Cluster CiPS^{M} (EXC114). Further support is obtained from the European Union's Horizon 2020 research and innovation program under grant agreement No. 642023 (ITN clickgene).

Keywords: epigenetic bases • click chemistry • 5-formylcytosine • genomic DNA • droplet digital PCR

[1] P. A. Jones, *Nat. Rev. Genet.* **2012**, *13*, 484-492.

[2] T. Carell, C. Brandmayr, A. Hienzsch, M. Müller, D. Pearson, V. Reiter, I. Thoma, P. Thumbs, M. Wagner, *Angew. Chem., Int. Ed.* **2012**, *51*, 7110-7131; *Angew. Chem.* **2012**, *124,* 7220-7242.

[3] a) S. Kriaucionis, N. Heintz, *Science* **2009**, *324*, 929-930; b) M. Tahiliani, K. P. Koh, Y. Shen, W. A. Pastor, H. Bandukwala, Y. Brudno, S. Agarwal, L. M. Iyer, D. R. Liu, L. Aravind, A. Rao, *Science* **2009**, *324*, 930-935.

[4] T. Pfaffeneder, B. Hackner, M. Truss, M. Münzel, M. Müller, C. Deiml, C. Hagemeier, T. Carell, *Angew. Chem., Int. Ed.* **2011**, *50*, 7008-7012; *Angew. Chem*. **2011**, *123*, 7146-7150.

[5] a) S. Ito, L. Shen, Q. Dai, S. C. Wu, L. B. Collins, J. A. Swenberg, C. He, Y. Zhang, *Science* **2011**, *333*, 1300-1303; b) Y.-F. He, B.-Z. Li, Z. Li, P. Liu, Y. Wang, Q. Tang, J. Ding, Y. Jia, Z. Chen, L. Li, Y. Sun, X. Li, Q. Dai, C.-X. Song, K. Zhang, C. He, G.-L. Xu, *Science* **2011**, *333*, 1303-1307.

[6] S. Schiesser, B. Hackner, T. Pfaffeneder, M. Müller, C. Hagemeier, M. Truss, T. Carell, *Angew. Chem., Int. Ed.* **2012**, *51*, 6516-6520; *Angew. Chem.* **2012**, *124*, 6622-6626.

[7] a) M. Münzel, D. Globisch, T. Brückl, M. Wagner, V. Welzmiller, S. Michalakis, M. Müller, M. Biel, T. Carell, *Angew. Chem., Int. Ed.* **2010**, *49*, 5375-5377; *Angew. Chem.* **2010**, *122*, 5503-5505; b) D. Globisch, M. Münzel, M. Müller, S. Michalakis, M. Wagner, S. Koch, T. Brückl, M. Biel, T. Carell, *PLoS ONE* **2010**, *5*, e15367; c) M. Wagner, J. Steinbacher, T. F. J. Kraus, S. Michalakis, B. Hackner, T. Pfaffeneder, A. Perera, M. Müller, A. Giese, H. A. Kretzschmar, T. Carell, *Angew. Chem., Int. Ed.* **2015**, *54*, 12511-12514; *Angew. Chem*. **2015**, *127*, 12691-12695.

[8] a) N. Plongthongkum, D. H. Diep, K. Zhang, *Nat. Rev. Genet.* **2014**, *15*, 647-661; b) M. J. Booth, E.-A. Raiber, S. Balasubramanian, *Chem. Rev.* **2015**, *115*, 2240-2254.

[9] a) W. A. Pastor, U. J. Pape, Y. Huang, H. R. Henderson, R. Lister, M. Ko, E. M. McLoughlin, Y. Brudno, S. Mahapatra, P. Kapranov, M. Tahiliani, G. Q. Daley, X. S. Liu, J. R. Ecker, P. M. Milos, S. Agarwal, A. Rao, *Nature* **2011**, *473*, 394-397; b) E.-A. Raiber, D. Beraldi, G. Ficz, H. Burgess, M. Branco, P. Murat, D. Oxley, M. Booth, W. Reik, S. Balasubramanian, *Genome Biol.* **2012**, *13*, R69; c) C.-X. Song, Keith E. Szulwach, Q. Dai, Y. Fu, S.-Q. Mao, L. Lin, C. Street, Y. Li, M. Poidevin, H. Wu, J. Gao, P. Liu, L. Li, G.-L. Xu, P. Jin, C. He, *Cell* **2013**, *153*, 678-691; d) B. Xia, D. Han, X. Lu, Z. Sun, A. Zhou, Q. Yin, H. Zeng, M. Liu, X. Jiang, W. Xie, C. He, C. Yi, *Nat. Methods* **2015**, *12*, 1047-1050.

[10] a) M. Yu, Gary C. Hon, Keith E. Szulwach, C.-X. Song, L. Zhang, A. Kim, X. Li, Q. Dai, Y. Shen, B. Park, J.-H. Min, P. Jin, B. Ren, C. He, *Cell* **2012**, *149*, 1368-1380; b) M. J. Booth, M. R. Branco, G. Ficz, D. Oxley, F. Krueger, W. Reik, S. Balasubramanian, *Science* **2012**, *336*, 934-937; c) M. J. Booth, G. Marsico, M. Bachman, D. Beraldi, S. Balasubramanian, *Nat. Chem.* **2014**, *6*, 435-440; d) X. Lu, C.-X. Song, K. Szulwach, Z. Wang, P. Weidenbacher, P. Jin, C. He, *J. Am. Chem. Soc.* **2013**, *135*, 9315-9317.

[11] a) A. Nomura, K. Sugizaki, H. Yanagisawa, A. Okamoto, *Chem. Commun.* **2011**, *47*, 8277-8279; b) J. Duprey, G. A. Bullen, Z.-Y. Zhao, D. M. Bassani, A. F. A. Peacock, J. Wilkie, J. H. R. Tucker, *ACS Chem. Bio.* **2016**, *11*, 717-721.

[12] a) P. M. E. Gramlich, S. Warncke, J. Gierlich, T. Carell, *Angew. Chem., Int. Ed.* **2008**, *47*, 3442-3444; *Angew. Chem*. **2008**, *120*, 3491-3493.; b) J. Willibald, J. Harder, K. Sparrer, K.-K. Conzelmann, T. Carell, *J. Am. Chem. Soc.* **2012**, *134*, 12330-12333.

[13] A. Maiti, A. C. Drohat, *J. Biol. Chem.* **2011**, *286*, 35334-35338.

[14] B. J. Hindson, K. D. Ness, D. A. Masquelier, P. Belgrader, N. J. Heredia, A. J. Makarewicz, I. J. Bright, M. Y. Lucero, A. L. Hiddessen, T. C. Legler, T. K. Kitano, M. R. Hodel, J. F. Petersen, P. W. Wyatt, E. R. Steenblock, P. H. Shah, L. J. Bousse, C. B. Troup, J. C. Mellen, D. K. Wittmann, N. G. Erndt, T. H. Cauley, R. T. Koehler, A. P. So, S. Dube, K. A. Rose, L. Montesclaros, S. Wang, D. P. Stumbo, S. P. Hodges, S. Romine, F. P. Milanovich, H. E. White, J. F. Regan, G. A. Karlin-Neumann, C. M. Hindson, S. Saxonov, B. W. Colston, *Anal. Chem.* **2011**, *83*, 8604-8610.

[15] M. Wendeler, L. Grinberg, X. Wang, P. E. Dawson, M. Baca, *Bioconjugate Chem.* **2014**, *25*, 93-101.

Table of contents

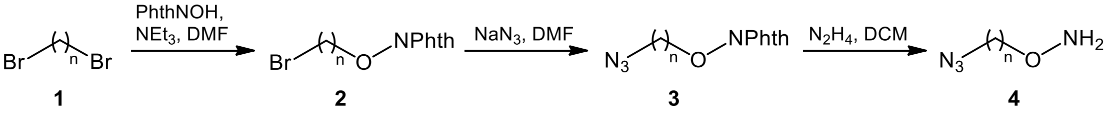
1. General methods of organic synthesis

Chemicals were purchased from *Sigma-Aldrich* and used without further purification. The solvents for organic synthesis were of reagent grade and purified by distillation. Solutions were concentrated *in vacuo* on a *Heidolph* rotary evaporator with a *Vario PC2001* diaphragm pump by *Vacuubrand*. All mixed solvent systems are reported as v/v solutions. All reactions were monitored by thin-layer chromatography (TLC), performed on *Merck* 60 (silica gel F_{254}) plates. Chromatographic purification of products was accomplished using flash column chromatography on silica gel (230-400 mesh) purchased from Merck.

^{1}H- and ^{13}C-NMR spectra were recorded in deuterated solvents on *Bruker ARX* 400 spectrometers and calibrated to the residual solvent peak. Chemical shifts (δ, ppm) are quoted relative to the residual solvent peak as internal standard and coupling constants (*J*) are corrected and quoted to the nearest 0.1 Hz. Multiplicities are abbreviated as follows: s = singlet, d = doublet, t = triplet, m = multiplet.

2. Synthesis of the hydroxylamine linker

O-(4-Azidobutyl)hydroxylamine 4



1,4-Dibromobutane **1** (5.9 mL, 49.4 mmol, 2.0 eq.) was added to a solution of *N*-hydroxyphthalimide (PhthNOH, 4.0 g, 24.5 mmol, 1.0 eq.) and triethylamine (7.5 mL, 53.6 mmol, 2.2 eq.) in anhydrous dimethylformamide. The mixture was stirred at room temperature for 24 h. The reaction was diluted with water, and the aqueous phase was extracted three times with ethyl acetate. The combined organic phases were dried over $MgSO_4$, filtered and concentrated to give the crude product **2** (5.02 g, 16.9 mmol, 0.69 eq.) as a white solid. The residue was dissolved in anhydrous dimethylformamide and sodium azide (1.32 g, 20.6 mmol, 0.85 eq.) was added. The mixture was stirred at room temperature for 2 h, diluted with water and extracted with ethyl acetate three times. The combined organic phases were dried over $MgSO_4$, filtered and concentrated. The crude was purified by flash chromatography on silica gel (*iso*-hexane/ethyl acetate 10:1→ 2:1) to give **3** (3.97 g, 15.2 mmol, 0.62 eq.) as a yellow oil. The oil was redissolved in hydrazine monohydrate (1.1 mL, 22.8 mmol, 0.93 eq.) and dichloromethane (10 mL). The mixture was stirred at room temperature for 24 h and then filtered. The solution was diluted with dichloromethane and washed with NaCl aq. three times. The combined organic phases were dried over $MgSO_4$, filtered and concentrated to give **4** (1.81 g, 13.9 mmol, 57% yield over three steps) as a colorless oil.

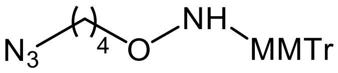
***R*f** = 0.42 (DCM/ MeOH 10:1).

^{1}H-NMR (400 MHz, CDCl$_3$): *δ* = 3.70 (t, $^3J_{H,H}$ = 5.6 Hz, 2H, O-C**H$_2$**-CH$_2$), 3.38 (t, $^3J_{H,H}$ = 6.4 Hz, 2H, C**H$_2$**-N$_3$), 1.68−1.64 (m, 4H, CH$_2$-C**H$_2$**-C**H$_2$**-CH$_2$).

^{13}C-NMR (100 MHz, CDCl$_3$): *δ* = 75.1 (-O-**C**H$_2$-CH$_2$), 51.2 (**C**H$_2$-N$_3$), 25.6 (**C**H$_2$), 25.5 (**C**H$_2$).

HRMS (ESI+): calculated for $C_4H_{11}ON_4^+$ [M+H]$^+$: 131.0927, found: 131.0928.

***O*-(4-Azidobutyl)-*N*-[(4-methoxyphenyl)diphenylmethyl]hydroxylamine** (5)



5

O-(4-Azidobutyl)hydroxylamine **4** (1.88 g, 14.4 mmol, 1.0 eq.) was dissolved in anhydrous dichloromethane (40 mL). 4-Monomethoxytritylchloride (MMTr-Cl, 4.91 g, 15.9 mmol, 1.1 eq.) and diisopropylethylamine (5.0 mL, 28.9 mmol, 2.0 eq.) was added to the mixture at 0°C. The reaction was stirred at room temperature for 2 h, diluted with dichloromethane, washed with saturated NaHCO$_3$, dried over MgSO$_4$, filtered and concentrated. The crude was purified by flash chromatography on silica gel (*iso*-hexanes/ethyl acetate 15:1 + 3% triethylamine) to give **5** (4.89 4g, 12.2 mmol, 84%) as a yellowish oil.

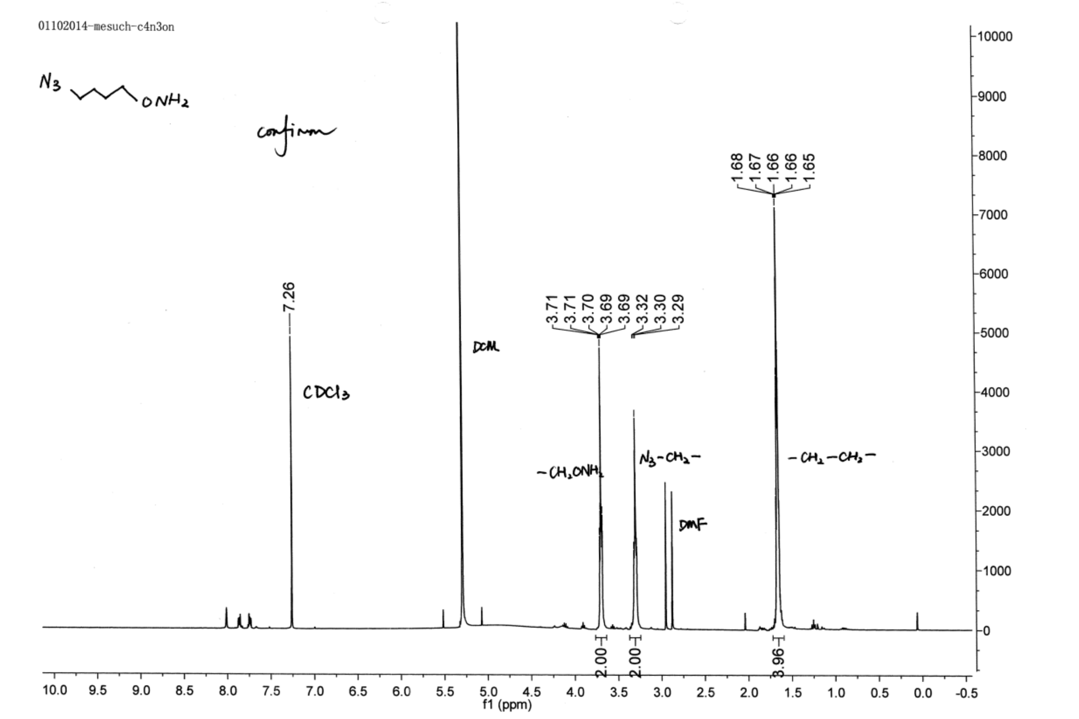
***R*f** = 0.64 (*iso*-hexane/ ethyl acetate 10:1 + 3% triethylamine).

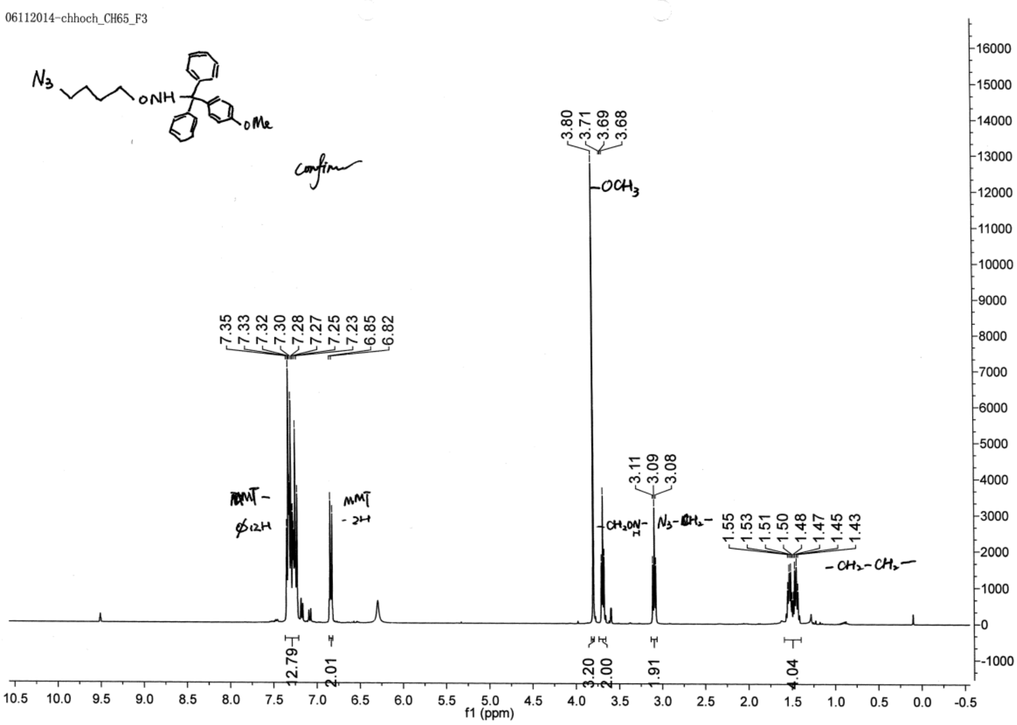
^{1}H-NMR (400 MHz, CDCl$_3$): *δ* = 7.31−7.18 (m, 12H, 12 × C$_{Ar}$**H**), 6.79 (d, $^3J_{H,H}$ = 8.8 Hz, 2H, 2 × CH$_3$-O-C-C**H**), 3.76 (s, 3H, O-C**H$_3$**), 3.65 (t, $^3J_{H,H}$ = 6.0 Hz, 2H, O-C**H$_2$**-CH$_2$), 3.05 (t, $^3J_{H,H}$ = 6.8 Hz, 2H, N$_3$-C**H$_2$**), 1.52−1.37 (m, 4H, N$_3$-CH$_2$-C**H$_2$**-C**H$_2$**-CH$_2$).

^{13}C-NMR (100 MHz, CDCl$_3$): *δ* = 158.3 (CH$_3$-O-**C**), 144.6 (2 × **C**$_{Ar}$), 136.6 (**C**$_{Ar}$), 130.2 (2 × **C**$_{Ar}$), 129.0 (4 × **C**$_{Ar}$), 127.6 (4 × **C**$_{Ar}$), 126.7 (2 × **C**$_{Ar}$), 112.9 (2 × **C**$_{Ar}$), 77.2 (O-NH-**C**), 73.2 (O-**C**H$_2$), 55.2 (O-**C**H$_3$), 51.1 (N$_3$-**C**H$_2$), 25.8 (N$_3$-CH$_2$-**C**H$_2$), 25.5 (O-CH$_2$-**C**H$_2$).

HRMS (ESI-): calculated for $C_{25}H_{27}O_4N_4^-$ [M+HCO$_2$]$^-$: 447.2038, found: 447.2040.

3. ^{1}H-NMR spectra of the linker





4. General methods for oligonucleotide synthesis

DNA Oligonucleotide synthesis was performed on an Applied Biosystems Incorporated 394 automated synthesizer. Phosphoramidites and solid supports columns were purchased from *Glen Research*, *Link Technology,* and *ChemGene Corporation*. Oligodeoxynucleotides were synthesized in a 1 μmol scale with standard DNA synthesis cycles (trityl off mode). Coupling time for modified nucleosides was extended to 10 min. The oligonucleotide was cleaved using conc. ammonium hydroxide aq. at 25 °C for 17 h. The aqueous solution was then collected and evaporated in a *SpeedVac* concentrator, and the pellet was redissolved in ddH_2O.

Analytical RP-HPLC was performed using a *Macherey-Nagel* Nucleodur 100-3 C18ec column on 2695 Separation Module equipped with a Waters Alliance 2996 Photodiode Array Detector using a flow of 0.5 mL/min. Semi-preparative RP-HPLC was performed using a *Macherey-Nagel* C18 column (5 mm, 9.4 × 250 mm) on *Waters* Breeze 2487 Dual λ Array Detector, 1525 Binary HPLC Pump. Conditions: Buffer A, 0.1 M TEAA (triethylammonium acetate) in water; buffer B, 0.1 M TEAA in 80% acetonitrile.

The purified fractions were concentrated and characterized by Matrix Assisted Laser Desorption Ionization Time of Flight (MALDI-TOF) on a Bruker Daltonics Autoflex II instrument. The concentration of the oligonucleotide solutions was calculated from the UV absorbance at 260 nm on a Nanodrop ND-1000 spectrophotometer. Extinction coefficients of the oligonucleotides at 260 nm were calculated by addition of the extinction coefficients of the individual nucleobases: dA 15.0 L/(mmol·cm), dC 7.1 L/(mmol·cm), dG 12.0 L/(mmol·cm), dT 8.4 L/(mmol·cm), mdC 7.8 L/(mmol·cm), hmdC 8.7 L/(mmol·cm), fdC 11.3 L/(mmol·cm), cadC 7.1 L/(mmol·cm). The 1,2,3-triazole and abasic monomer are transparent at 260 nm.

Click reaction on the solid support with azide linkers and deprotection

After solid phase synthesis (0.2 μmol scale, approx. 50% yield for 13 mer, calculated as 0.1 μmol), the solid support was suspended in dimethyl sulfoxide (80 μL) and acetonitrile (25 μL). To the mixture, $CuSO_4$ aq. (100 mM, 50 μL, 5.0 μmol, 50 eq.), sodium ascorbate aq. (500 mM, 20 μL, 10 μmol, 100 eq.), *N*,*N*-diisopropylethylamine solution in acetonitrile (200 mM, 75 μL, 15 μmol, 150 eq.), solution of **5** in dimethyl sulfoxide (100 mM, 50 μL, 5.0 μmol, 50 eq.) were added. The reaction was conducted at 25 °C for 24 h. Afterwards, the solid support was washed with dimethyl sulfoxide, dilute $NaHCO_3$ aq., acetonitrile, ether and air-dried to a powder. The solid phase was then cleaved with conc. aqueous NH_3 at 25 °C for 17 h, purified by HPLC and concentrated. The removal of the MMTr protection group on the hydroxylamine was carried out by dissolving the obtained oligonucleotide in acetic acid aq. (20%, 200 μL) at 25 °C for 30 min, precipitated by addition of sodium acetate solution (3 M, 60 μL) and EtOH (1040 μL), and then purified again with HPLC.

Schiff base formation between fdC-oligonucleotides and probe strands

A mixture of fdC containing oligonucleotides (**T1**,**2**), or T/C/mC/hmC/caC/Abasic containing oligonucleotides (**T3-8**) in control experiments (15 μM, 20 μL, 0.3 nmol, 1 eq.), probe strands (15 μM, 20 μL, 0.3 nmol, 1 eq.), NaCl aq. (1 M, 15 μL), NaOAc aq. (pH = 6.0, 100 mM, 15 μL) and ddH_2O (80 μL) was prepared to make a final volume of 150 μL (oligonucleotide working concentration 2 μM). The mixture was heated to 85 °C for 5 min then slowly cooled down to 25 °C in 3 h. A first aliquot (15 μL) was taken and quenched before 5.4 μL of 4-methoxyaniline solution (250 mM, ddH_2O/DMSO, v/v 9/1, acidified to pH = 5.5 with acetate acid) was added to give a catalyst working concentration of 10 mM. The reaction was conducted at 25 °C, 500 rpm for 24 h. Aliquots (15 μL) were taken at 1, 2, 4, 6, 8, 12, 24 h and quenched by addition of loading buffer. All the samples were heated at 85 °C for 3 min followed by PAGE assay as mentioned above.

When using 1,4-benzenediamine as the catalyst, a 10 mM stock solution of 10 mM in 0.5% acetic acid aq. was prepared.

Melting point experiments

Melting profiles were measured on a JASCO V-650 spectrometer using quartz glass cuvettes with 10 mm path length. The samples contained 100 mM NaCl, 10 mM NaOAc buffer pH 6.0 and 1 μM of each strand in a final volume of 200 μL. The measurement was repeated three times with independent sample. Before the measurement, the oligonucleotides were hybridized by slowly cooling down the samples from 85 °C to room temperature. The solutions were covered with silicon oil and tightly plugged. Absorbance was recorded in the forward and reverse direction at temperatures from 25 °C (or 15 °C) to 85 °C with a slope of 1 °C/min. T_M values were calculated as the zero-crossing of 2^{nd} derivate of the 339 nm background corrected change in hyperchromicity at 260 nm.

Table S1. Synthesized oligonucleotides in this study. (Letters in bold and italic stand for 2’-*O*-propargyl nucleosides or epigenetic modifications.)

Entry	Description	5'-------------3'	Calc.	Exptl.
T1	ODN-fC	GTAATG**fC**GCTAGG	4040.9	4036.2
T2	fC-shift	GTAAT**fC**CGCTAGG	4000.9	3999.6
T3	fC-T	GTAATG**T**GCTAGG	4027.7	4024.7
T4	fC-C	GTAATG**C**GCTAGG	4012.7	4008.6
T5	fC-mC	GTAATG**mC**GCTAGG	4026.7	4021.6
T6	fC-hmC	GTAATG**hmC**GCTAGG	4042.7	4039.6
T7	fC-caC	GTAATG**caC**GCTAGG	4056.7	4054.9
T8	fC-Abasic	GTAATG**AP**GCTAGG	3919.6	3916.4
P1	Probe-1-alkyne	CC***U***AGCGCATTAC	3932.7	3930.5
	Probe-1-MMTr	CC***U***AGCGCATTAC	4335.2	4335.1
	Probe-1-ONH$_2$	CC***U****AGCGCATTAC	4084.8[a]	4084.2
P2	Probe-2-MMTr	CC***U***ATCGCATTAC	4310.2	4310.3
	Probe-2-ONH$_2$	CC***U****ATCGCATTAC	4059.8[a]	4064.3

[a] contains one sodium ion

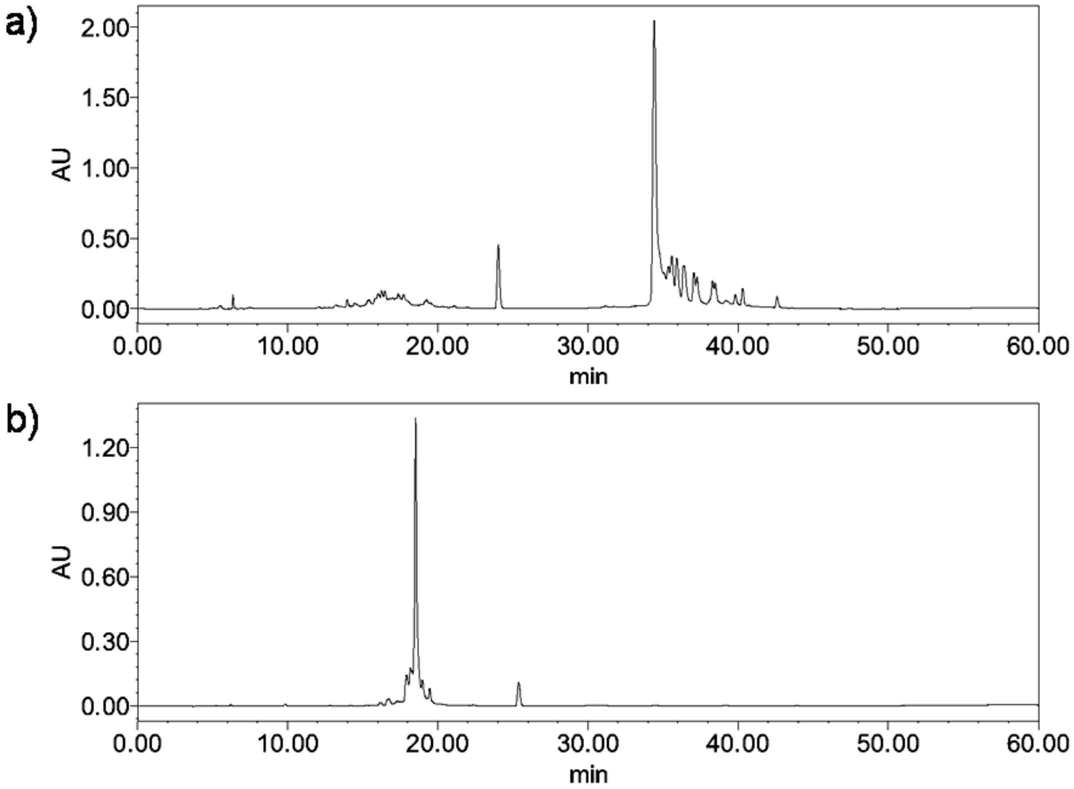


Figure S1. Typical HPLC trace of crude product a) **P1-MMTr** and b) **P1-ONH$_2$**. Conditions: buffer A, 0.1 M TEAA; buffer B, 0.1 M TEAA in 80% acetonitrile, linear gradient from 0% to 60% B over 45 min. Retation time: (a) 34.4 min, (b) 18.5 min. AU = arbitrary unit.

5. Crosslinking studies with the synthesized strands

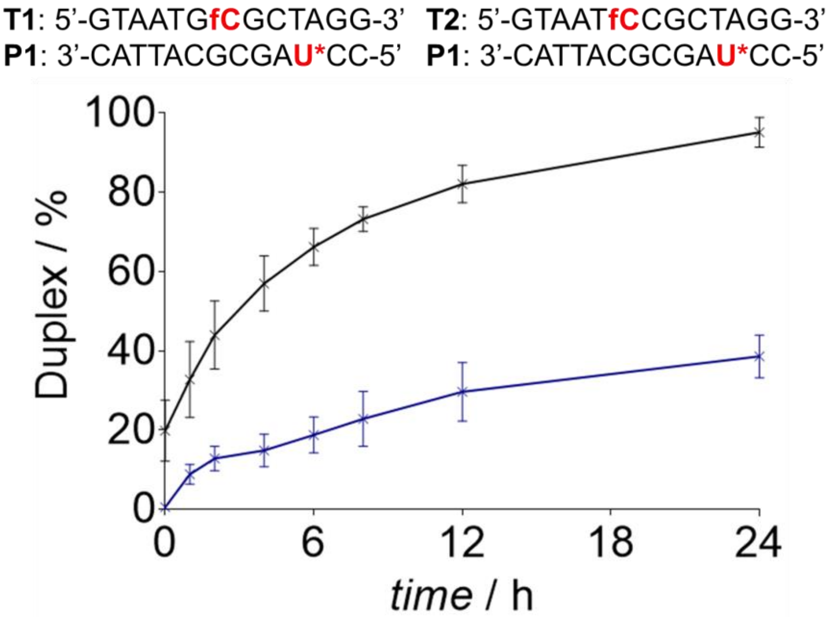


Figure S2. Quantification of the DNA duplex formation during the reaction using the catalyst 4-methoxyaniline. Black line: duplex formation between **T1** and **P1**, blue line: duplex formation between **T2** and **P1**. Error bars represent the standard error of the mean calculated from three replicates.

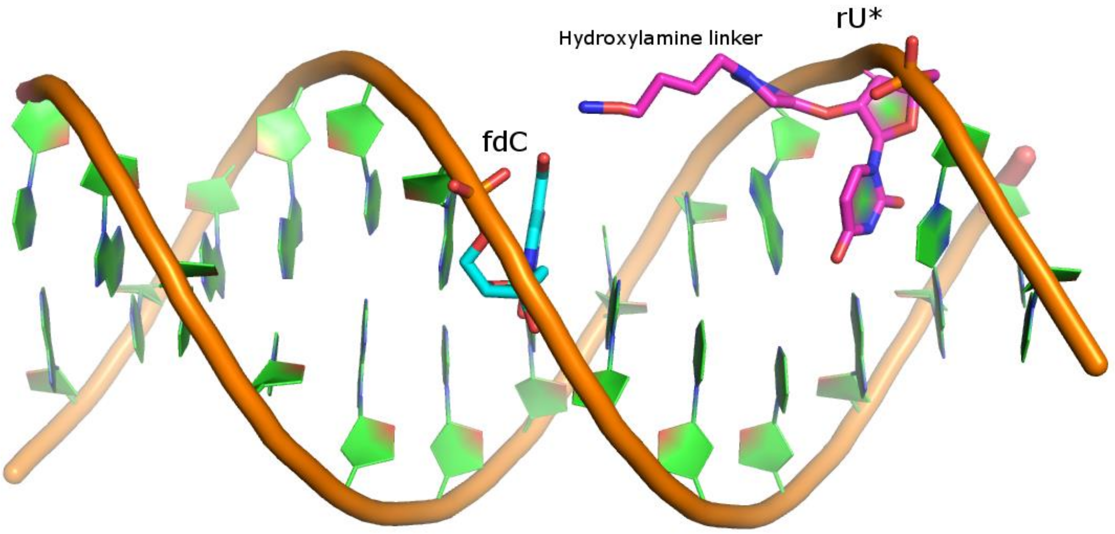
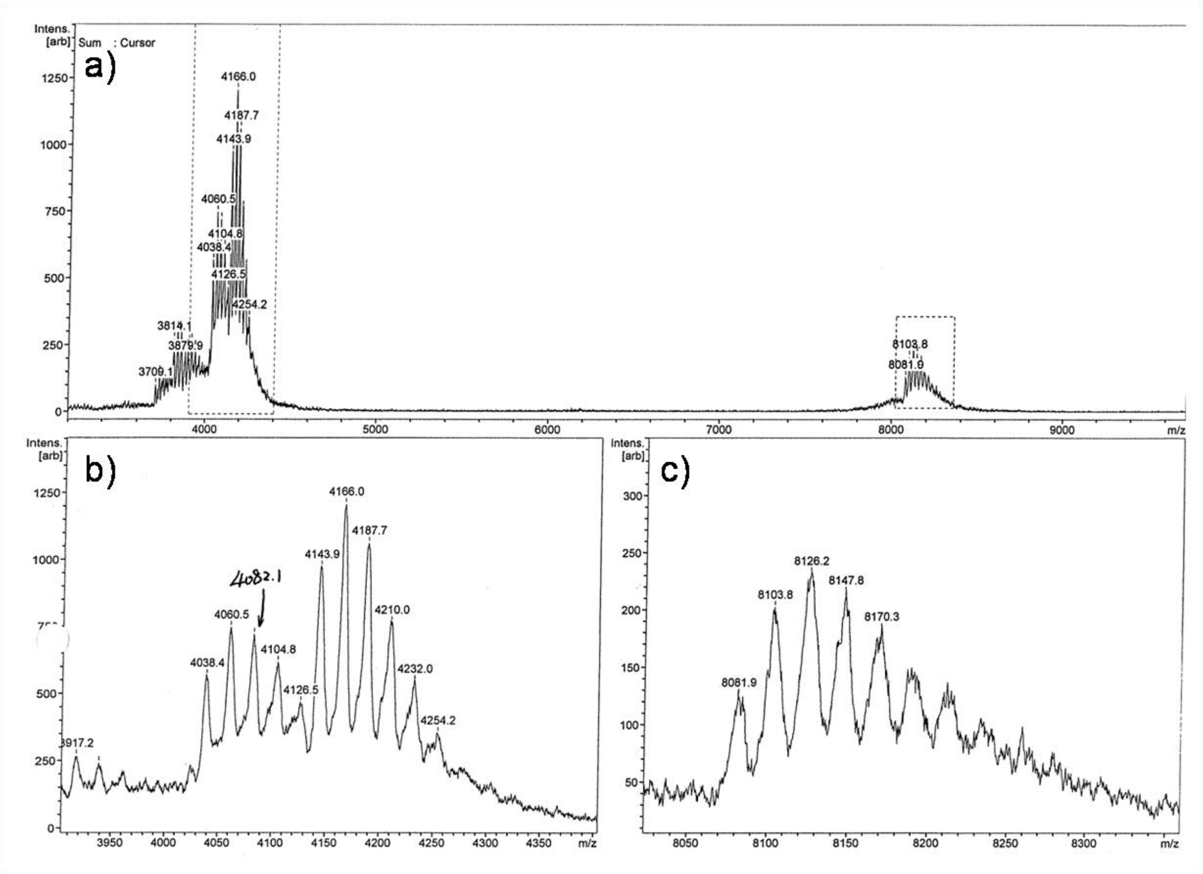


Figure S3. Model representation of a duplex showing the position of the hydroxylamine linker relative to the fdC on the complementary strand.



No.	5'-------------3'	Calc.	Exptl.
P1	CCU*AGCGCATTAC	4061.8	4060.5
T1	GTAATGfCGCTAGG	4040.9	4038.4
T1	GTAATGfCGCTAGG	4146.0[a]	4143.9
T1:P1		8084.7	8081.9

[a] conjugate with 4-methoxyaniline

Figure S4. MALDI-TOF mass spectrum of the crosslinked duplex **T1:P1** and single strands. a) Overall MALDI-TOF spectrum; b) peaks corresponding to single strands **P1** and **P1-** 4-methoxyaniline conjugate; c) peaks corresponding to linked duplex **T1:P1.** Conditions: 10 μM oligonucleotides, 100 mM NaCl, 10 mM NaOAc buffer pH 6.0.

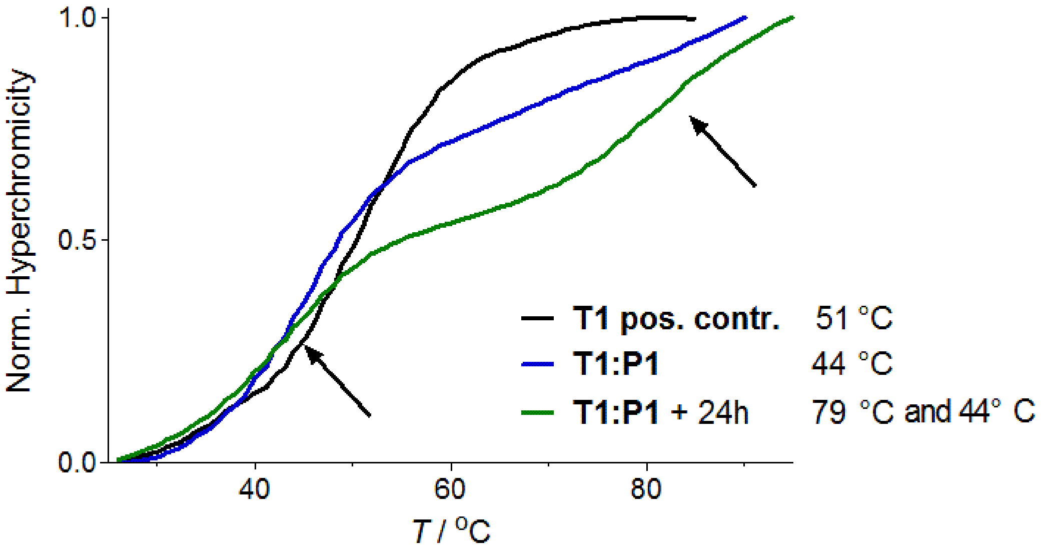


Figure S5. Melting curves of duplex **T1**:**P1** after reannealing or after 24 h incubation without catalyst compared with duplex **T1** and its counter strand (positive control). Conditions: 1 µM oligonucleotides, 100 mM NaCl, 10 mM NaOAc buffer pH 6.0, the final volume of 200 µL.

6. Experimental details of the genomic fdC profiling study

Cell culture and genomic DNA isolation

J1 wild type stem cells (strain 129/SvJae),[1] Dnmt TKO (J1, strain 129/SvJae),[2] Tdg$^{+/-}$ (E14, strain 129/Ola) and the Tdg$^{-/-}$ cell line (E14, strain 129/Ola),[3] were routinely maintained on gelatinized plates in DMEM (Sigma-Aldrich) supplemented with 10% FBS (PAN Biotech), 1x MEM-nonessential amino acids (NEAA), 0.2 mM L-alanyl-L-glutamine, 1x penicillin-streptomycin, 0.1 mM ß-mercaptoethanol (all from Sigma-Aldrich), 1000 U/ml mouse recombinant LIF (ORF Genetics), 1 μM PD 0325901 and 3 μM CHIR 99021 (2i; both from Axon Medchem). In these conditions, global genomic mC levels are very low and its oxidized derivatives are even lower, as we described previously.[4] For the experiments, the cultures were passaged twice (over five days), in DMEM supplemented with FBS and LIF, but lacking 2i. With this strategy, primed mESC cultures were obtained and oxidized cytosine derivatives reached reproducibly higher and stable levels.[4] In case of the experiment using J1 wild type and Dnmt TKO cells, the cultures were passaged every second day over a period of six days.

Mouse embryonic stem cells were lysed directly in the plates with RLT-buffer (Qiagen). The lysates were homogenized with a TissueLyser MM400 (Retsch) for 1 min at 30 Hz and centrifuged for 5 min at 21000 xg. Then genomic DNA was isolated using the Zymo Quick gDNA Midi Kit according to the manufacturer’s instruction. The concentration was measured using a Nanodrop ND-1000 (Peqlab).

Probe crosslinking

The gDNA solution obtained above (1.2 μg), the fdC probe (**P**) (1 μM, 2 μL), NaH_2PO_4-Na_2HPO_4 buffer (200 mM, pH = 6.0, 2 μL), NaCl aq. (1.5 M, 2μL), and ddH_2O were mixed to a final volume of 18 μL. The mixture was heated to 95 °C for 3 min, and then cooled down rapidly to 25 °C. 1,4-Benzenediamine aq. (10 mM, 2 μL) was added and the reaction vial shaken (300 rpm) for 6 h at 25 °C. First, the mixture was neutralized with Na_2HPO_4 aq. (200 mM, 40 μL), and then purification with *NEB Monarch* PCR DNA Cleanup Kit using the binding buffer (120 μL) and eluting with the elution buffer (Tris-EDTA) (30 μL). The eluted solution was quantified with the Nanodrop and 22-32 ng/μL was obtained. UV spectra confirmed the main peak centered at 260 nm.

Ligation

The crosslinked gDNA solution (300 ng), reporter strand (**R**) (20 nM, 1 μL), Ampligase reaction buffer (10×, 2 μL), Ampligase from *Epicenter* (5 U/μL, 2 μL, 10 U) and ddH$_2$O were mixed to a final volume of 20 μL. The mixture was heated to 95 °C for 3 min, and then to 94 °C for 1 min, 60 °C for 1 h and back to 94 °C for 10 cycles. Then, the reaction mixture was diluted with Tris-HCl buffer (200 mM, pH = 7.6, 50 μL) before purification with *NEB Monarch* PCR DNA Cleanup Kit using the binding buffer (140 μL) and eluting with the elution buffer (10 μL). The eluted solution was quantified with the Nanodrop, obtaining 20-32 ng/μL. UV spectra confirmed the main peak centered at 260 nm.

Droplet digital PCR

ddPCR experiments were performed on a *Bio-Rad* QX100 ddPCR System. For one reaction, gDNA (6 ng), four primers (18 μM each, 1 μL), two TaqMan probes (5 μM each, 1 μL), digital PCR Supermix for Probes (no dUTP, 2×, 10 μL), and ddH$_2$O were mixed to a final volume of 20 μL with primer working concentration of 900 nM and TaqMan probe working concentration of 250 nM.

PCR cycles were conducted on a *Bio-Rad* T100 Thermal cycler. PCR cycle: 95°C for 10 min, 94°C for 30 sec and specific annealing temperature (64°C) for 1 min for 35 or 40 cycles, then 98°C for 10 min and cooled down to 12°C. A temperature ramp of 2°C/s was used. Droplet generation and counting were conducted according to the manufacturer's instructions, i.e. reaction mixture prepared as above (20 μL) and ddPCR droplet generate oil (70 μL) were used per reaction. The accounted droplet number was retained in 10000-18000. FAM for detection amplicon was set to channel 1; HEX for reference amplicon was set to channel 2.

Each percentage value represents the averages and standard deviations from the mean of at least two technical replicates and two biological replicates. LC-MS quantification were conducted according to the previous report.[5]

References

[1] E. Li, T. H. Bestor, R. Jaenisch, *Cell* **1992**, *69*, 915-926.

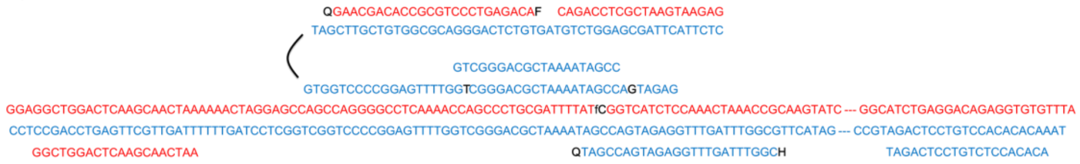
[2] A. Tsumura, T. Hayakawa, Y. Kumaki, S. Takebayashi, M. Sakaue, C. Matsuoka, K. Shimotohno, F. Ishikawa, E. Li, H. R. Ueda, J. Nakayama, M. Okano, *Genes Cells* **2006**, *11*, 805-814.

[3] D. Cortazar, C. Kunz, J. Selfridge, T. Lettieri, Y. Saito, E. MacDougall, A. Wirz, D. Schuermann, A. L. Jacobs, F. Siegrist, R. Steinacher, J. Jiricny, A. Bird, P. Schar, *Nature* **2011**, *470*, 419-423.

[4] T. Pfaffeneder, F. Spada, M. Wagner, C. Brandmayr, S. K. Laube, D. Eisen, M. Truss, J. Steinbacher, B. Hackner, O. Kotljarova, D. Schuermann, S. Michalakis, O. Kosmatchev, S. Schiesser, B. Steigenberger, N. Raddaoui, G. Kashiwazaki, U. Muller, C. G. Spruijt, M. Vermeulen, H. Leonhardt, P. Schar, M. Muller, T. Carell, *Nat Chem Biol* **2014**, *10*, 574-581.

[5] M. Wagner, J. Steinbacher, T. F. J. Kraus, S. Michalakis, B. Hackner, T. Pfaffeneder, A. Perera, M. Müller, A. Giese, H. A. Kretzschmar, T. Carell, *Angew. Chem., Int. Ed.* **2015**, *54*, 12511-12514.

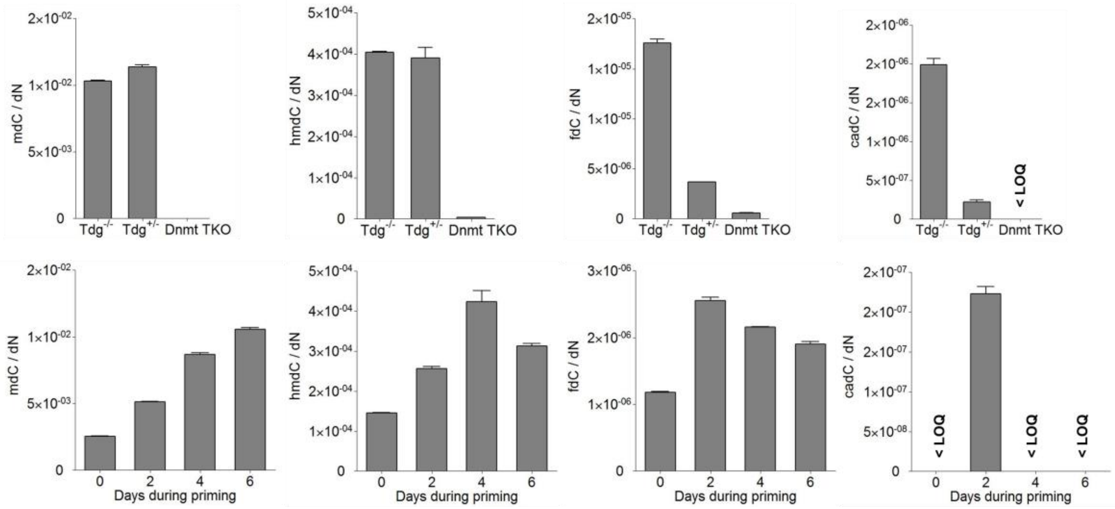
a)


QGAACGACACCGCGTCCCTGAGACAF
CAGACCTCGCTAAGTAAGAG
TAGCTTGCTGTGGCGCAGGGACTCTGTGATGTCTGGAGCGATTCATTCTC
GTCGGGACGCTAAAATAGCC
GTGGTCCCCGGAGTTTTGGTCGGGACGCTAAAATAGCCAGTAGAG
GGAGGCTGGACTCAAGCAACTAAAAAACTAGGAGCCAGCCAGGGGCCTCAAAACCAGCCCTGCGATTTTATfCGGTCATCTCCAAACTAAACCGCAAGTATC --- GGCATCTGAGGACAGAGGTGTGTTTA
CCTCCGACCTGAGTTCGTTGATTTTTTGATCCTCGGTCGGTCCCCGGAGTTTTGGTCGGGACGCTAAAATAGCCAGTAGAGGTTTGATTTGGCGTTCATAG --- CCGTAGACTCCTGTCCACACACAAAT
GGCTGGACTCAAGCAACTAA
QTAGCCAGTAGAGGTTTGATTTGGCH
TAGACTCCTGTCTCCACACA

b)

Entry	Description	5' ---------- 3'	mer
P2	probe for locus 1	GAGATG*ACCGATAAAATCGCAGGGC	25
R1	report strand for locus 1	pTGGTTTTGAGGCCCCTGGTGTAGCTTGCTGTGGCG CAGGGACTCTGTGATGTCTGGAGCGATTCATTCTC	70
locus 1	detection forward primer	CCGATAAAATCGCAGGGCTG	20
	detection reverse primer	GAGAATGAATCGCTCCAGAC	20
	detection TaqMan probe	F-ACAGAGTCCCTGCGCCACAGCAAG-Q	24
	reference forward primer	GGCTGGACTCAAGCAACTAA	20
	reference reverse primer	ACACACCTCTGTCCTCAGAT	20
	reference TaqMan probe	H-CGGTTTAGTTTGGAGATGACCGAT-Q	24
P3	probe for locus 2	CA*AGCGAAGCAGGGCAAATGGCGA	24
R2	report strand for locus 2	pTCTCGAACCTCTGCCAGCCTAGCTTGCTGTGGCGC AGGGACTCTGTGATGTCTGGAGCGATTCATTCTC	69
locus 2	detection forward primer	AAGCAGGGCAAATGGCGA	19
	detection reverse primer	GAGAATGAATCGCTCCAGAC	20
	detection TaqMan probe	F-ACAGAGTCCCTGCGCCACAGCAAG-Q	24
	reference forward primer	CCAGGGAGCATCTGTGAAAA	20
	reference reverse primer	AAAAGCCCATCTGGGAAACA	20
	reference TaqMan probe	H-CCCCTTCAGACGCAAGCGAAGCAG-Q	24

Figure S6. Sequence detail of the detection strategy: a) Illustration of the detection strategy with sequence details for locus 1, i.e. 30,020,539th position on chromosome 16 (MM9); b) synthesized and purchased oligonucleotides for locus 1 and 2. Bold and red letters represent nucleoside modifications or functional group: p, phosphate group at 5’ terminus; F, FAM; H, HEX; Q, BHQ-1. MALDI-TOF: **P2**: calc.7951.5, found 7948.5, **P3** calc. 7672.5, found 7671.0, contain one sodium ion.



	mdC		hmdC		fdC		cadC	
	pro dN	STABW%	pro dN	STABW%	pro dN	STABW%	pro dN	STABW%
Tdg-/-	1.0E-02	1.1	4.0E-04	1.1	1.8E-05	3.6	2.0E-06	7.3
Tdg+/-	1.1E-02	2.0	3.9E-04	11.3	3.7E-06	0.8	2.2E-07	22.3
Dnmt TKO	3.0E-06	28.0	3.8E-06	18.8	6.0E-07	6.8	*	*
WT0	2.5E-03	1.7	1.5E-04	0.6	1.2E-06	1.9	*	*
WT2	5.1E-03	1.6	2.6E-04	3.6	2.6E-06	3.3	2.2E-07	6.8
WT4	8.7E-03	2.4	4.2E-04	11.2	2.2E-06	0.8	*	*
WT6	1.1E-02	1.7	3.1E-04	3.1	1.9E-06	3.4	*	*

Figure S7. Global 5mC, 5hmC, fdC, and 5caC quantification using LC-MS: Tdg$^{-/-}$, Tdg$^{+/-}$, and Dnmt TKO mES cells and 0, 2, 4, 6 days during priming of wild-type mES cells. *: <LOQ, below the limit of quantification.

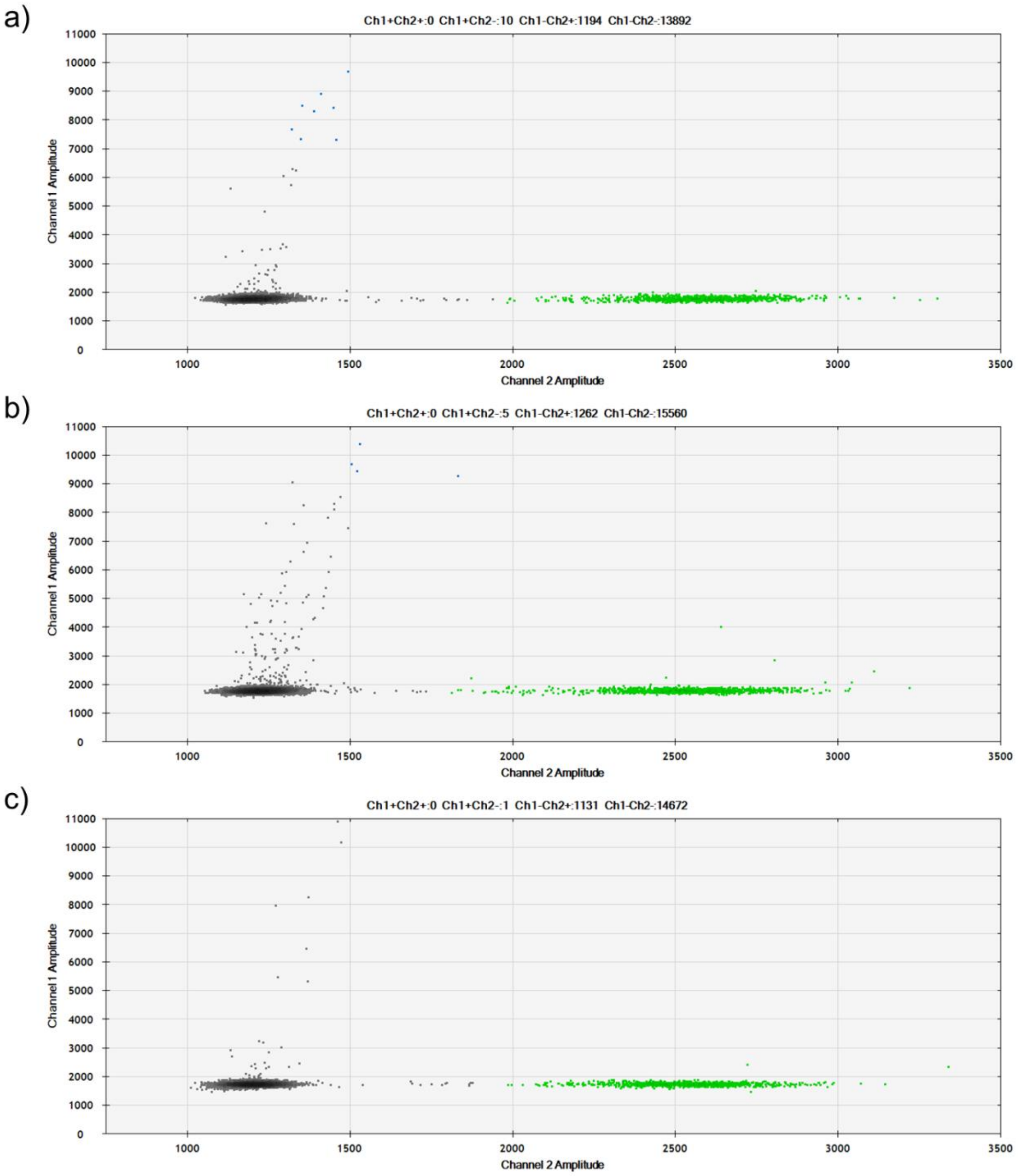


Figure S8. 2-D plot of droplet fluorescence for negative control of locus 1: a) **P3**, instead of **P2**, was used; b) no reporter stand **R1**; c) no Ampligase.

7. Quantification modeling

The encapsulation maximum of one target amplicon in one droplet to generate a positive or negative signal is the ideal scenario for our situation. If a droplet contains more than one detection amplicon, for example, one contains one fdC and one cytosine, it will show a positive signal, and the negative cytosine signal vanishes.

The probability for two or more detection amplicons to get into one droplet can be calculated according to the Poisson distribution, i.e. a discrete random variable X complies the Poisson distribution with parameter λ > 0, if, for k = 0, 1, 2, …, the probability mass function of X is given by:

$$f\ (k;\lambda) = \Pr\ (X = k) = \frac{\lambda^k e^{-\lambda}}{k!}$$

where e is Euler's number and *k*! is the factorial of *k*.

Table S2 Poisson distribution probabilities of genome copies in the droplet.

input ng	λ	*k*				
		1	2	3	4	5
3	0.05	4.8%	0.1%	0.0%	0.0%	0.0%
6	0.10	9.0%	0.5%	0.0%	0.0%	0.0%
9	0.15	12.9%	1.0%	0.0%	0.0%	0.0%
10	0.17	14.1%	1.2%	0.1%	0.0%	0.0%
15	0.25	19.5%	2.4%	0.2%	0.0%	0.0%
20	0.33	23.9%	4.0%	0.4%	0.0%	0.0%
30	0.50	30.3%	7.6%	1.3%	0.2%	0.0%
40	0.67	34.2%	11.4%	2.5%	0.4%	0.1%

For example, the mass of a mouse genome is approximately 3.0 pg (3.0×10^{-12} g). If 30 ng for a 20 µL reaction is used, 10,000 genomes will be distributed into 20,000 droplets. So, λ equals to 10,000 / 20,000 = 0.50. Let X = 1, then *f* (1; 0.50) = 0.303; let X = 2, then *f* (2; 0.50) = 0.076. This means 30.3% of the droplets, instead 50% of the droplet, contain a single copy while 7.6% of the droplets contain two copies. Extensive distribution probabilities are listed in Table S2. If less than 9 ng is settled in a 20 µL reaction, the probability to have two copies inside one droplet will be lower than 1%. For the ease of calculation, 6 ng gDNA is used for each reaction of 20 µL, corresponding to ca. 90 copy/µL.

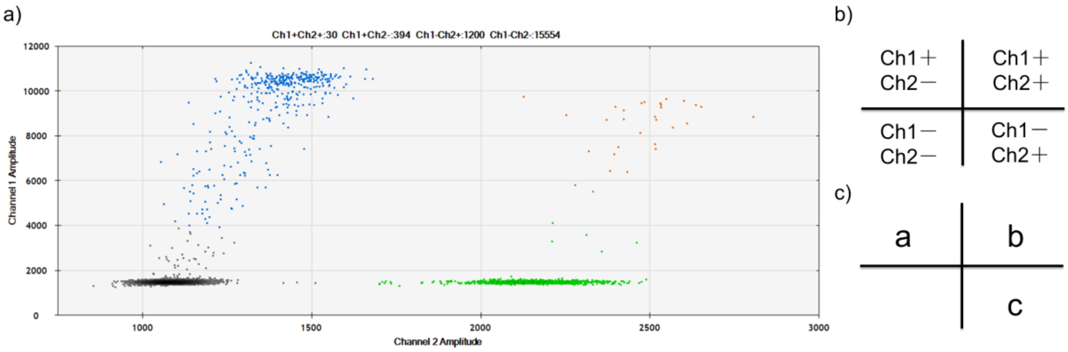


Figure S9. ddPCR output and modeling: a) 2-D plot of droplet fluorescence for illustration; b) clusters separation in four quadrants; c) algebraic simplification of counting numbers of the clusters.

As shown in Figure S8, Ch1+Ch2+ (yellow) refers to droplets with both positive signals; Ch1+Ch2- (blue) refers to droplets with only detection (report strand) signal; Ch1-Ch2+ (green) refers to droplets with only reference (gDNA) signal; Ch1-Ch2- (black) refers to droplets without target locus and ligated product; AD refers to all the droplets accepted; resolution refers to the separation of the clusters.

In principle, Ch1+Ch2+ shows the droplets that contain fdC sites in the target locus; Ch1+Ch2- indicates false-positive signals due to unspecific amplification and the dissociated ligated products; Ch1-Ch2+ shows the droplets containing only the target gDNA.

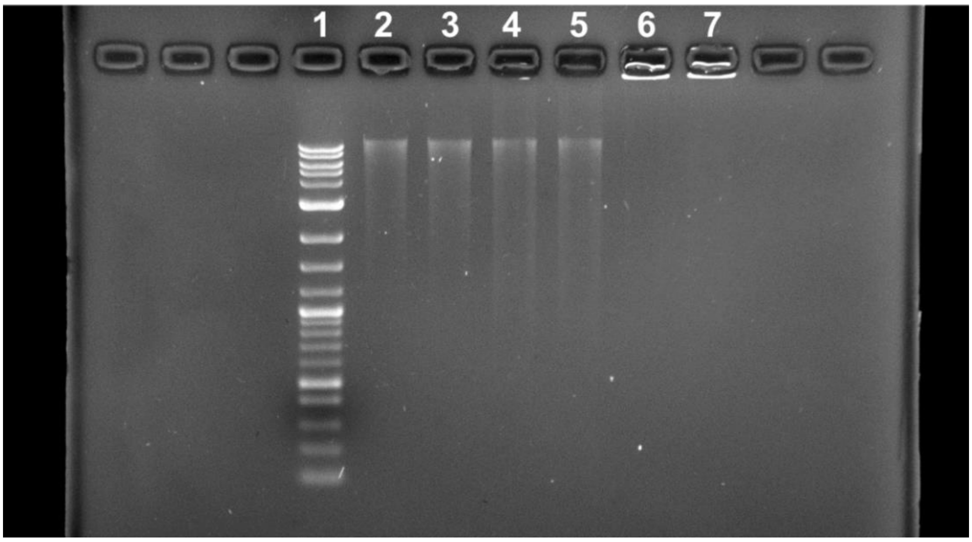


Figure S10. Agarose gel showing gDNA degradation: Line 1, log 2 marker; line 2,3, gDNA (150 ng) after crosslinking; line 4,5, gDNA (150 ng) after ligation cycle, 95°C for 3min, then 10 cycles of 94°C for 1min and 60°C for 1h; line 6,7, gDNA (150 ng).

Without considering the dissociation of the ligated products, the unreacted probe which remained in the system will cause unspecific amplification, i.e. Ch1+Ch2- signals. Catalyst, acid buffer, and ligation cycles will cause gDNA degradation (Figure S9, giving more Ch1+Ch2- false-negative signals. However, Ch1+Ch2- and Ch1+Ch2- / Ch1-Ch2- resolution do not play a role in the mathematical modelling that we used.

Assuming that all fdC at the target site is converted to the reporter strand via crosslinking and ligation, the yield is 100%. Assume that there are less than 150 copies in 1 µL so that the Poisson distribution is exclusive in our model.

Let a = Ch1+Ch2-, b = Ch1+Ch2+, c = Ch1-Ch2+, (Figure S8) A = Accepted droplet for the experiment entry, respectively, a’, b’, c’, and A’ for the control, i.e. TET knockout cell line.

Let η = fdC content of the target site.

Then,

$$\eta = \frac{\mathrm{a} + \mathrm{b} - \frac{A}{A'}(a' + b')}{b + c + (a - \frac{A}{A'}a')/\eta}$$

where $\mathrm{a} - \frac{\mathrm{A}}{\mathrm{A}'}a'$ refers to the degraded gDNA copy containing fdC at the target site that does not show in Ch2, $(\mathrm{a} - \frac{A}{A'}a')/\eta$ refers to all the degraded gDNA copies.

So,

$$\eta = \frac{(\mathrm{b} - \frac{A}{A'}\mathrm{b}')}{b + c}$$

Herein, in this ideal model, without considering the dissociation of the ligated products, η is independent of a, a’, and c’, i.e. genome degradation and unspecific ligation do not affect fdC percentage. Ch1+Ch2- only be resulted from the dissociation of the ligated products. Also, as shown in Figure S6 a-c, b’ can be omitted. Simplified η'

$$\eta' = \frac{\mathrm{a} + \mathrm{b}}{b + c}$$

is calculated to indicate relative abundance of fdC at the target site.

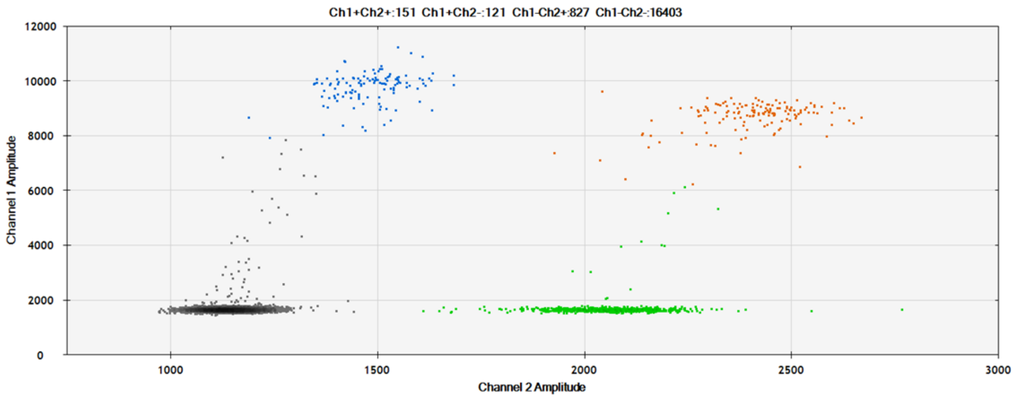
In reality, neither the fdC probe covers all the target sites nor the reporter strand ligates to all the target-linked probe. Therefore, only relative quantification is possible in our model.

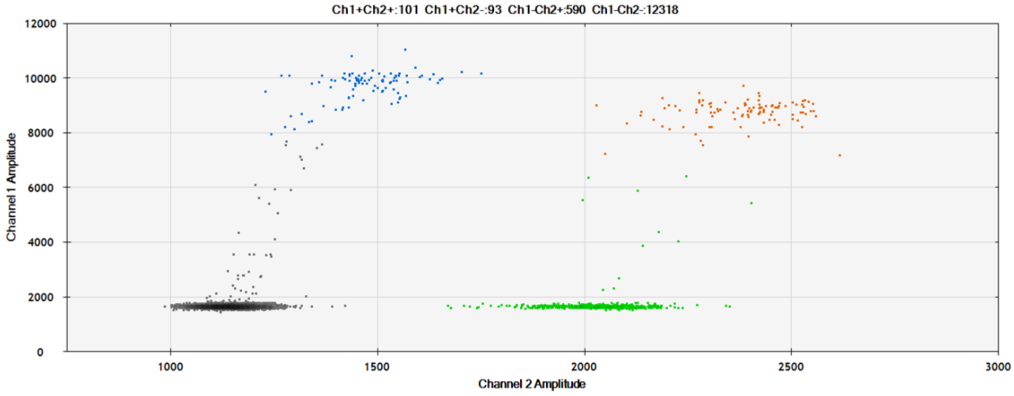
8. Droplet digital PCR data

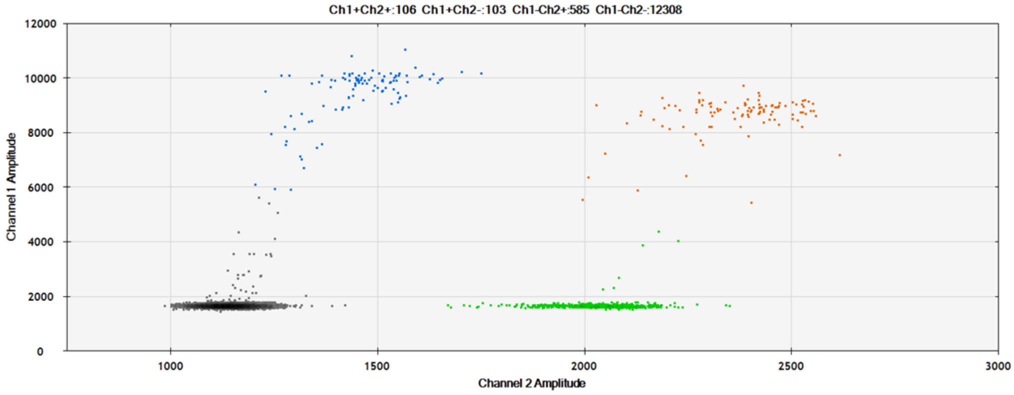
Raw data of fdC detection in Tdg$^{-/-}$, Tdg$^{+/-}$, Dnmt TKO cells for locus 1. (AD: accepted droplets)

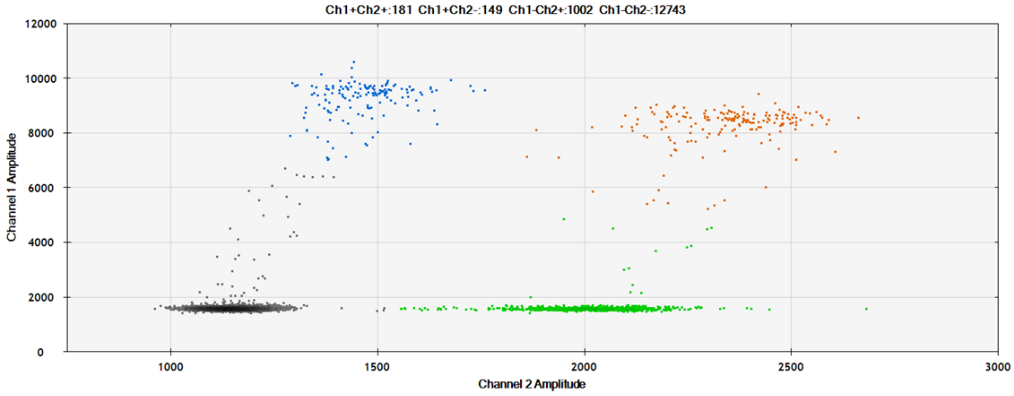
	Ch1	Ch2	1+2+	1+2-	1-2+	1-2-	AD	η	Average
Tdg-/-	17.6	67.6	151	121	827	16399	17502	27.8%	28.5%
Tdg-/-	17.6	63.7	101	93	590	12318	14896	28.1%	
Tdg-/-	18.9	63.7	106	103	585	12308	13102	30.2%	
Tdg-/-	28.8	103	181	149	1002	12743	14075	27.9%	
Tdg+/-	11.6	74.0	54	56	676	11085	11878	15.1%	15.7%
Tdg+/-	11.3	67.2	54	70	698	12722	13544	16.5%	
Tdg+/-	10.7	69.6	52	83	806	13994	14935	15.7%	
Tdg+/-	10.8	72.0	56	66	738	12490	13350	15.4%	
Dnmt TKO	4.0	115.0	10	29	1049	10327	11415	3.7%	5.2%
Dnmt TKO	3.1	61.1	11	17	625	11913	12566	4.4%	
Dnmt TKO	4.1	58.0	14	30	620	12508	13174	6.9%	
Dnmt TKO	2.8	60.6	8	21	600	11484	12119	4.8%	
Dnmt TKO	5.7	96.0	35	28	986	11998	13047	6.2%	

Locus 1 Tdg$^{-/-}$ mES cell sample

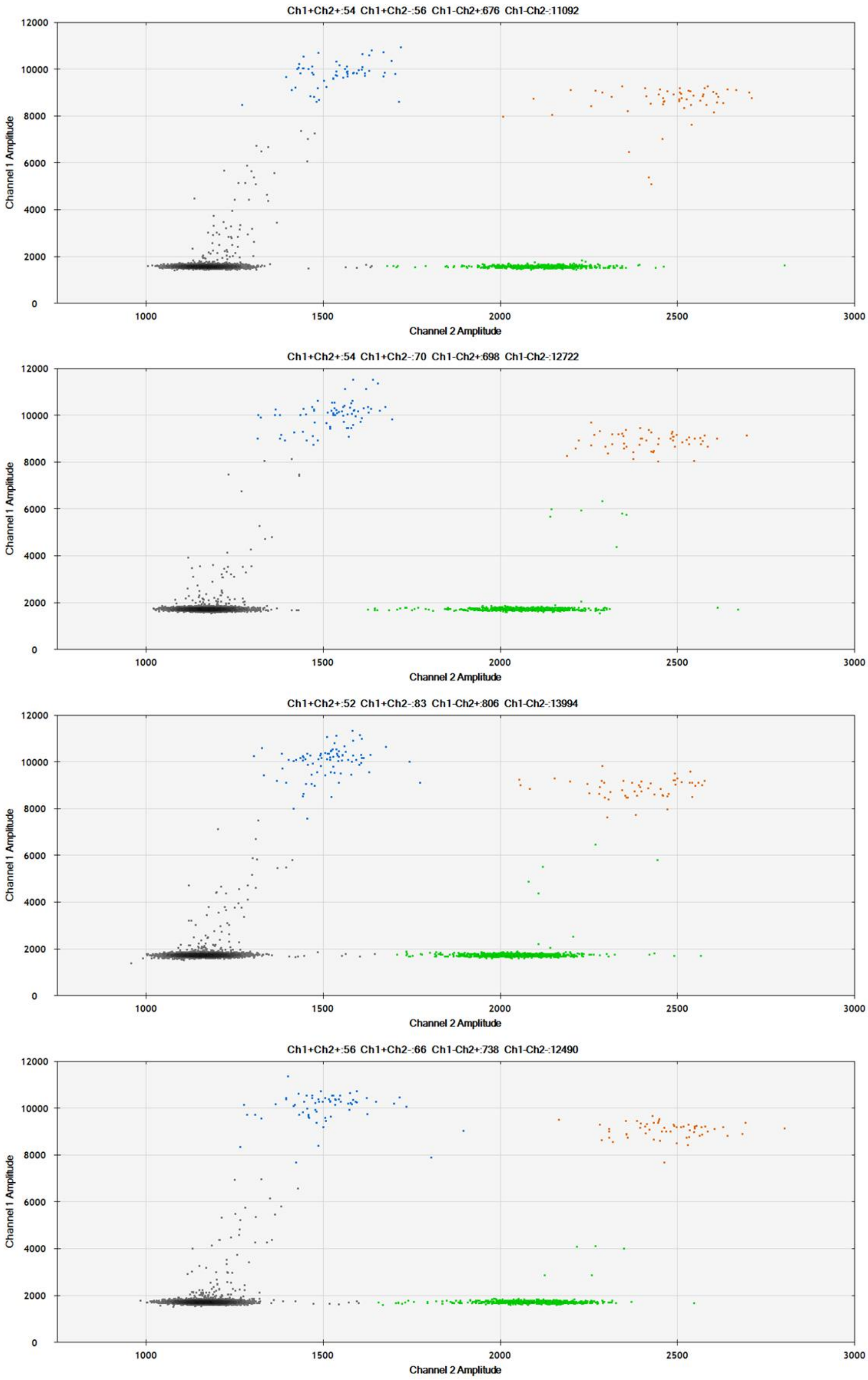




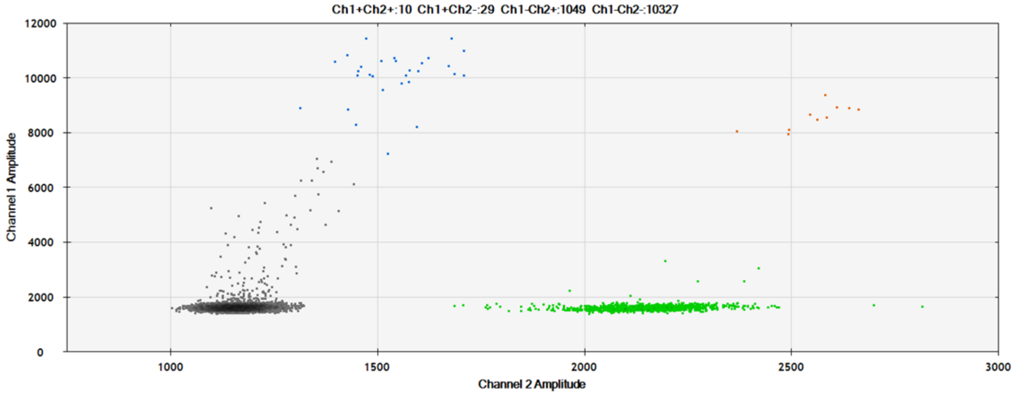


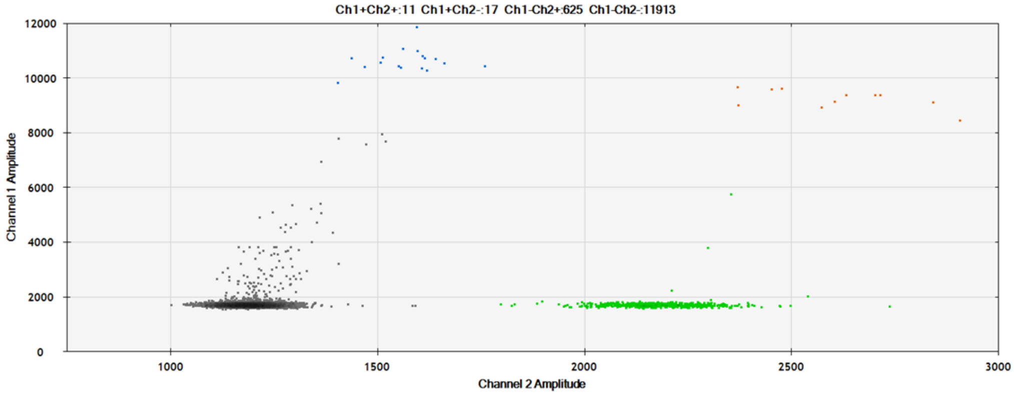


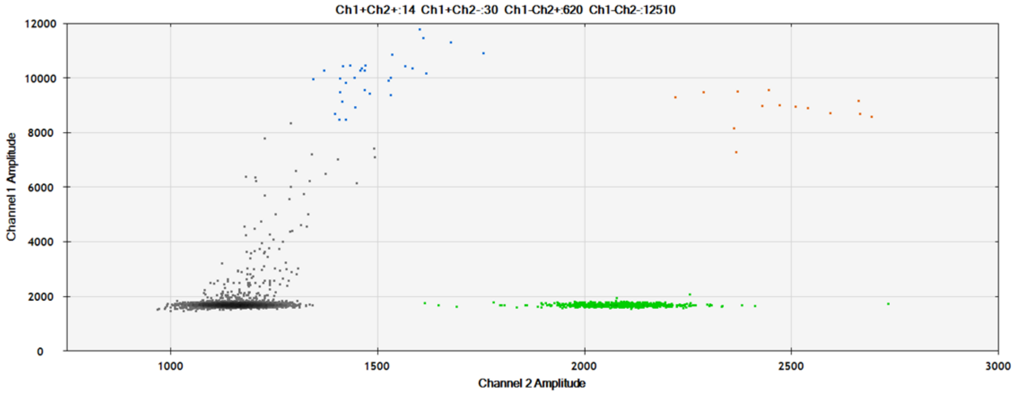
Locus 1 Tdg$^{+/–}$ mES cell sample.

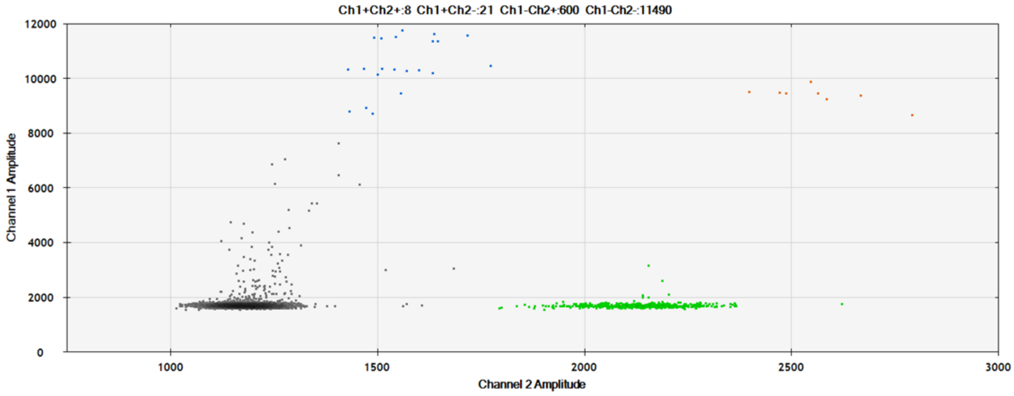


Locus 1 Dnmt TKO mES cell sample.









Locus 1 Raw data of fdC detection in wild-type cells during priming.

	Ch1	Ch2	1+2+	1+2-	1-2+	1-2-	AD	η	Average
WT0	5.4	108	20	34	1017	10787	11858	5.21%	6.06%
WT0	6.8	110	30	58	1336	13841	15265	6.44%	
WT0	6.6	112	26	53	1255	12738	14072	6.17%	
WT0	7.8	108	23	58	1241	13060	14382	6.41%	
WT2	9.7	89.2	40	98	1193	15552	16883	11.19%	10.90%
WT2	10.2	88.7	35	105	1249	15948	17337	10.90%	
WT2	10.8	93.0	35	96	1168	14476	15771	10.89%	
WT2	10.2	96.0	54	94	1279	15637	17064	11.10%	
WT2	6.3	56.2	19	65	744	15582	16410	11.01%	
WT2	6.0	57.3	27	54	759	15780	16620	10.31%	
WT4	6.2	68.0	19	30	501	8702	9252	9.42%	8.76%
WT4	5.4	68.4	26	46	811	13942	14821	8.60%	
WT4	4.5	59.0	14	24	477	9449	9964	7.74%	
WT4	5.8	64.5	23	43	688	12579	13333	9.28%	
WT6	11.1	117	42	67	1081	10691	11881	9.71%	8.67%
WT6	8.2	114	28	41	888	9000	9957	7.53%	
WT6	9.8	120	32	41	820	7909	8802	8.57%	
WT6	8.5	114	43	57	1086	11010	12196	8.86%	

Locus 2 Raw data of fdC detection in Tdg$^{-/-}$, Tdg$^{+/-}$cells

Well		Ch1	Ch2	1+2+	1+2-	1-2+	1-2-	AD	η	Average
H04	Tdg-/-	10.5	52	55	92	650	15671	16468	20.9%	19.8%
E06	Tdg-/-	20.8	115	88	133	1084	11285	12590	18.9%	
F06	Tdg-/-	22.5	114	98	173	1226	12805	14302	20.5%	
G06	Tdg-/-	20.1	109	94	149	1178	12927	14348	19.1%	
H06	Tdg-/-	21.8	116	90	159	1179	12141	13569	19.6%	
B09	Tdg+/-	5.5	66	28	49	884	15696	16657	8.4%	9.2%
A10	Tdg+/-	5.9	59	13	29	400	8002	8444	10.2%	
E01	Tdg+/-	8.3	96	43	78	1299	15749	17169	9.0%	
G01	Tdg+/-	8.1	93	39	73	1205	15070	16387	9.0%	