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.

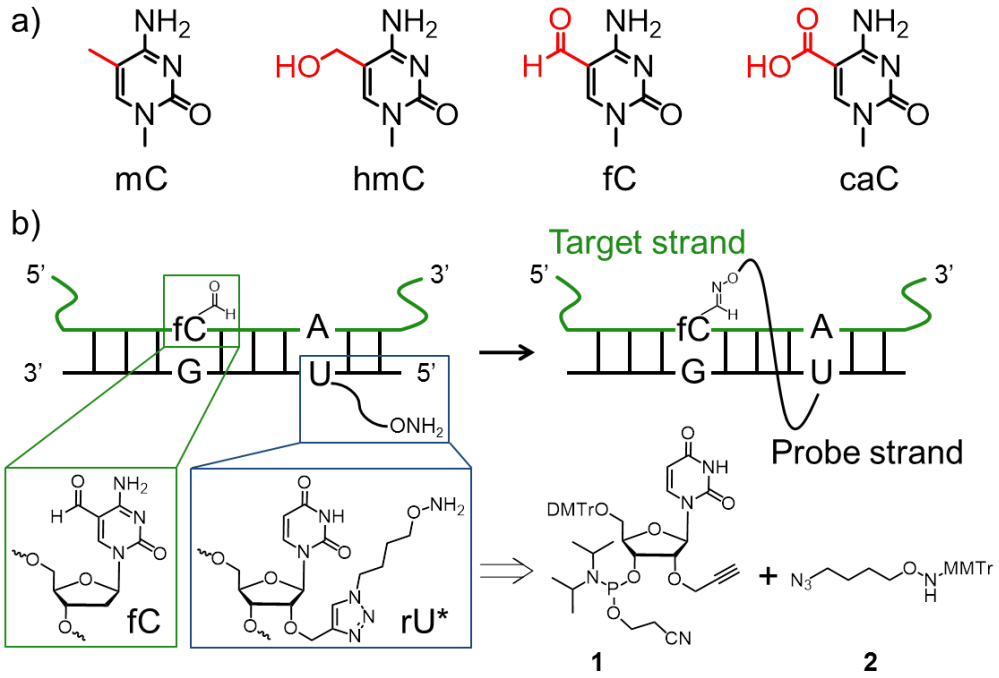

Figure 1. a) Structures of cytosine epigenetic modifications; b) Schematic representation of the fdC detection strategy, and used building blocks including the click chemistry-based assembly of the rU* probe molecule.

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).

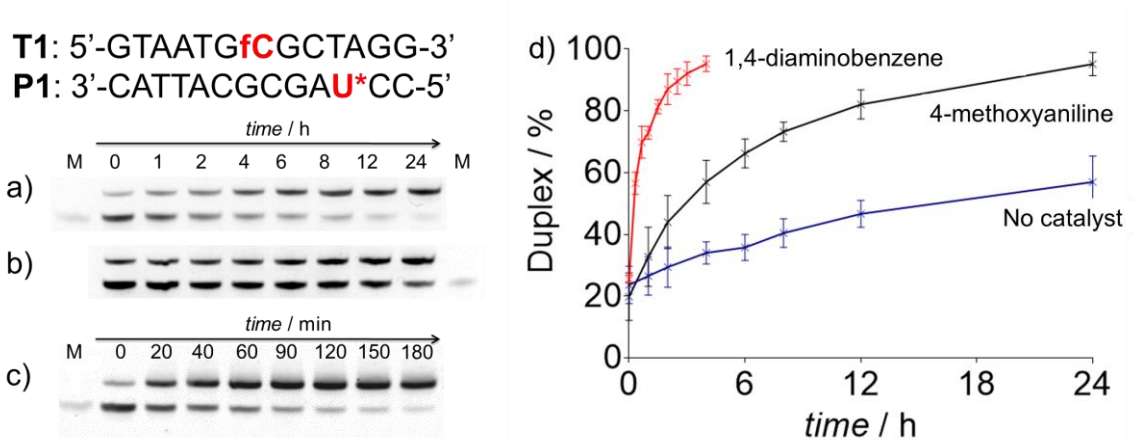

Figure 2. Denaturing PAGE gel showing the duplex formation between T1 and P1 at 25°C: a) with the catalyst 4-methoxyaniline; b) without a catalyst; c) with the catalyst 1,4-diaminobenzene; d) Quantification of the DNA duplex formation during the reaction. Black: catalyst 4-methoxyaniline, blue: no catalyst, 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.

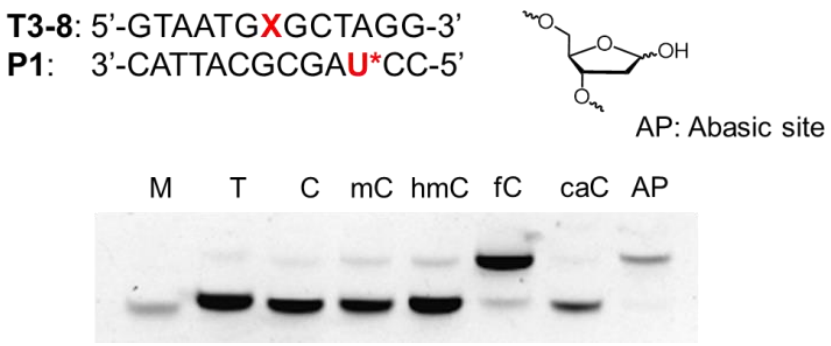

Figure 3. Denaturing PAGE gel showing duplex formation of T3-8 and P1 at 25°C after 24 h.

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.

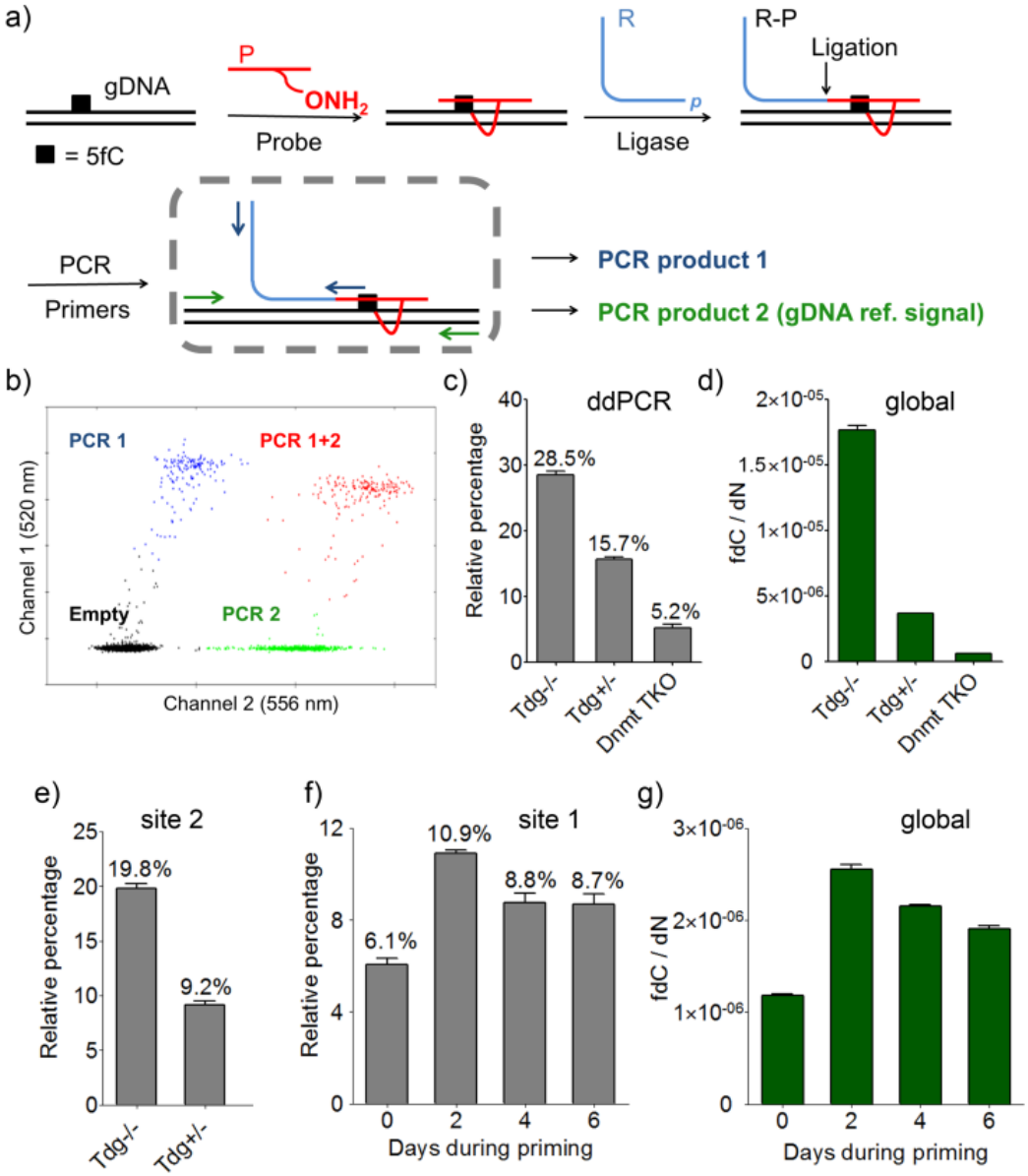

Figure 4. The fdC sequencing method: a) Schematic representation of the method, black line: gDNA; red segments: fdC probe; blue segments: reporter strands; arrows: PCR primer pairs; b) Typical 2-D plot of droplet fluorescence; c) Cluster ratios for position 1 in Tdg-/-, Tdg+/- and Dnmt TKO mES cells; 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.