Bioorganic Chemistry

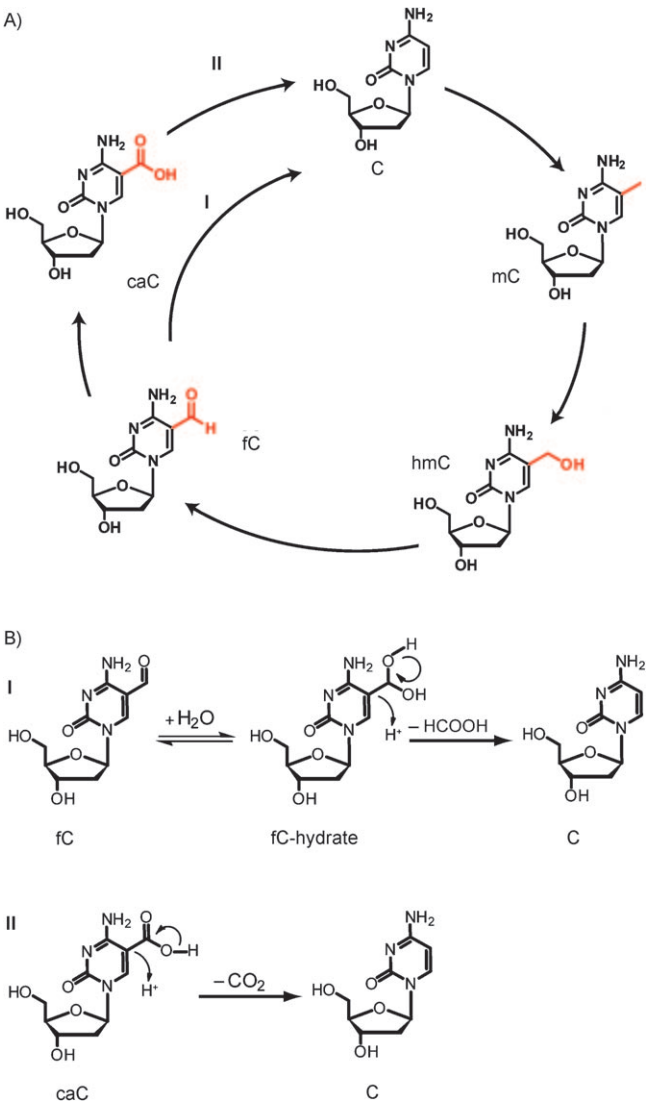
DOI: 10.1002/anie.201103899

The Discovery of 5-Formylcytosine in Embryonic Stem Cell DNA**

Toni Pfaffeneder, Benjamin Hackner, Matthias Truß, Martin Münzel, Markus Müller, Christian A. Deiml, Christian Hagemeier, and Thomas Carell**

Cellular development requires the silencing and activation of specific gene sequences in a well-orchestrated fashion. Transcriptional gene silencing is associated with the clustered methylation of cytosine bases (C) in CpG units of promoters. The methylation occurs at position C5 of cytosine to give 5-methylcytosine (mC) with the help of special DNA methyltransferases (DNMT).[1] The DNA methylome is significantly reprogrammed at various stages during early development,[2] during the development of primordial germ cells,[2c,3] or later in a locus-specific way at postdevelopmental stages.[4] Decreasing levels of mC can be established passively by successive rounds of DNA replication in the absence of methyltransferases. Active demethylation, in contrast, is proposed to be a process in which the mC bases are directly converted back into unmodified cytosines in the genome.[5] The recent discovery that mC can be further oxidized to hydroxymethylcytosine (hmC) with the help of TET enzymes[6] has led to the idea that hmC is connected to epigenetic reprogramming,[7] maybe as an intermediate in an, as yet controversial, active demethylation process.[4,5,8] Indeed recent data suggest that active demethylation in postdevelopmental phases may proceed through deamination of hmC to give 5-hydroxymethyluridine (hmU), which is then removed from the genome with the help of the base excision repair (BER) system.[9] Chemically, an attractive alternative mechanism for a more global active demethylation could be envisioned through further oxidation of hmC to give either 5-formylcytosine (fC) or 5-carboxylcytosine (caC) followed by elimination of a formyl or carboxyl group, respectively (Scheme 1).[5a,10] Although such an oxidative active demethylation pathway with hmC as the starting point has been frequently postulated,[5a,10] none of the further oxidized bases (fC, caC) have so far been detected.[10a]

To examine the question of whether hmC is the only oxidized base present in genomic DNA or if other higher oxidized species may be present as well, we performed an HPLC-MS study using synthetic fC and caC material as



Scheme 1. A) Putative cycle of methylation and oxidative demethylation of cytidine derivatives. B) Details of the demethylation reaction via vinyl carbanions.[11] **I**: Deformylation of fC to C. **II**: decarboxylation of caC to C.

[*] M. Sc. T. Pfaffeneder,[+] M. Sc. B. Hackner,[+] Dipl.-Chem. M. Münzel, Dr. M. Müller, Dipl.-Biochem. C. A. Deiml, Prof. Dr. T. Carell
CIPSM, Fakultät für Chemie und Pharmazie
Ludwig-Maximilians-Universität München
Butenandtstrasse 5–13, 81377 München (Germany)
E-mail: thomas.carell@lmu.de
Homepage: http://www.carellgroup.de
Dr. M. Truß,[+] Prof. Dr. C. Hagemeier
Charité Universitätsklinikum, Otto-Heubner-Centrum für Kinder- und Jugendmedizin, Klinik für Allgemeine Pädiatrie, Labor für Pädiatrische Molekularbiologie
Ziegelstrasse 5–9, 10098 Berlin (Germany)
E-mail: matthias.truss@charite.de

[+] These authors contributed equally to this work.

[**] We thank the Deutsche Forschungsgemeinschaft (DFG) for financial support through the SFB program (SFB 749, 646, and TRR54) and by grant CA275 8/4. Further support was obtained from the Excellence Cluster (Center for Integrative Protein Science, CiPS[M]) and NGFNplus (01GS0870).

Supporting information for this article is available on the WWW under http://dx.doi.org/10.1002/anie.201103899.

standards.[10a,12] Specifically, we used DNA isolated from mouse embrionic stem cells and embyoid bodies (mEBs) after two and three days of differentiation for the study. In mES cells, CpG-rich gene promoters are actively maintained in a hypomethylated state and differentiation of mES cells induces a wave of gene-specific de novo methylation that involves repression of TET1 and TET2 expression as well as reduction of global hmC and an increase in the global mC levels.[7d] In addition, we analyzed DNA isolated from nerve cell tissue, which features the highest levels of hmC in mice. Indeed, hippocampus and cortex tissue have hmC levels of up to approximately 0.7 %/G,[12a] which is about twofold higher than the hmC values measured in mES cells (0.39 %/G). The chromatogram obtained of a fully digested DNA sample from mES cells shows the signals for the four canonical bases A, C, G, and T plus the signal for mC (Figure 3A).[10a,12a] If detection is performed by mass spectrometry, the correct high-resolution mass (m/z values) for these five compounds and additionally the mass signal for hmC can be clearly detected. To our surprise, we detected in addition to these six signals one more signal of a compound that eluted with a retention time very close to T (Figure 1 A). This signal was initially only detectable in the DNA material isolated from mES cells. The new signal could not be directly detected in DNA isolated from mEBs, but unknown fragment ions were seen in further MS^n studies. The unknown compound detected in the mES cell DNA had a high-resolution mass signal (m/z_{found} = 256.0929) which is in excellent agreement with the calculated exact mass for fC (m/z_{calcd} = 256.0928; Figure 1 B). To unequivocally prove that the signal is generated by the presence of fC we synthesized the fC compound independently, as reported previously by us,[10a] and co-injected a small amount of the synthetic material into the DNA digest obtained from the mES cell DNA. Indeed, the synthetic material eluted with the same retention time (see the Supporting Information). Finally MS^n-fragmentation experiments were performed, which are highly compound specific. In these studies, the fC compound is fragmented directly in the mass spectrometer to give characteristic fragment ions. (Figure 1 C, D) The MS/MS data obtained from the putative fC compound isolated from mES cell DNA were found to be identical with literature data[13] and with the MS/MS data obtained from the authentic synthetic fC material (data not shown). In addition, the obtained MS^n data were identical with the unknown fragment ions detected in the mEB cell DNA. These data prove that the newly discovered compound

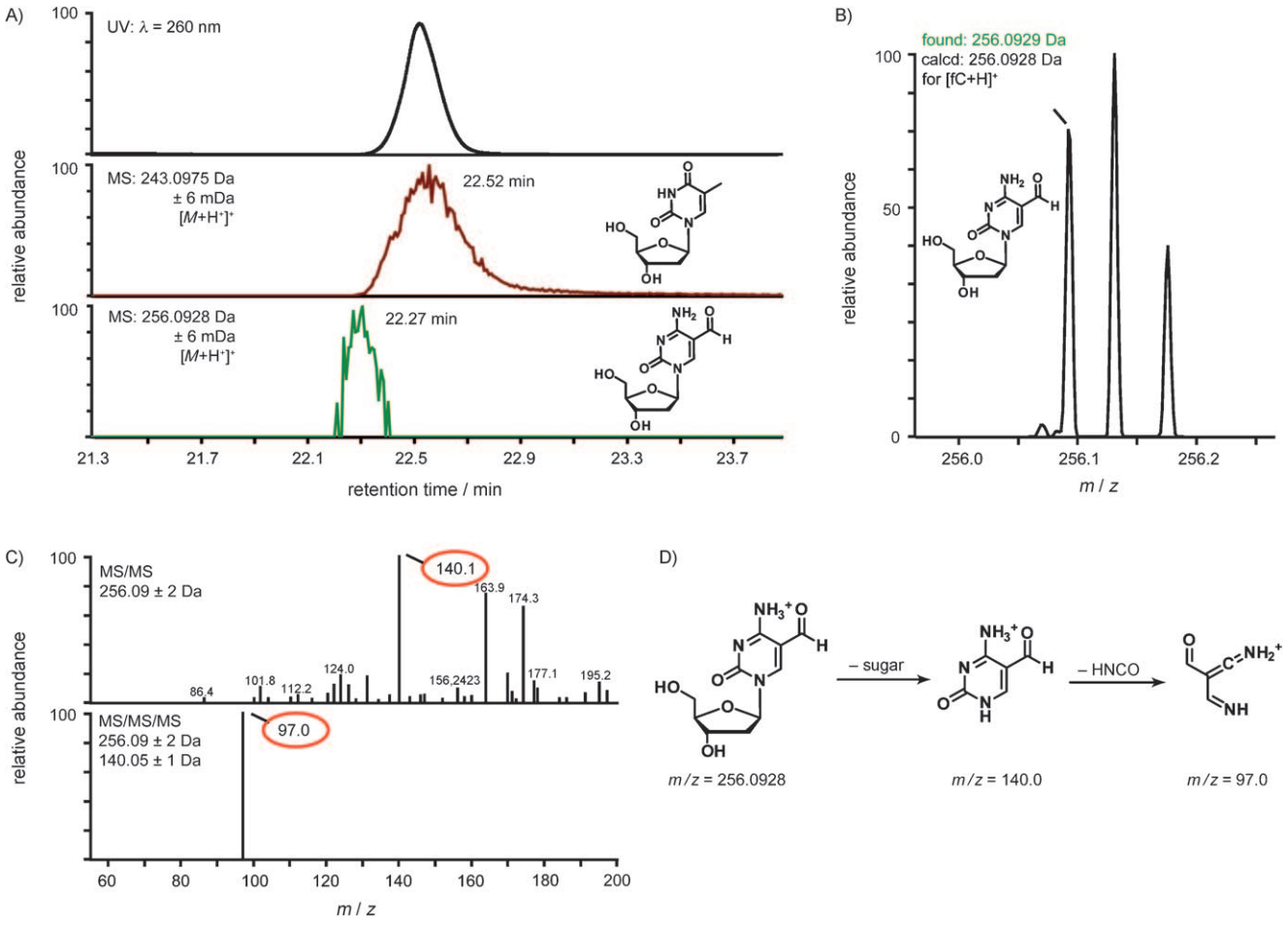


Figure 1. A) HPLC trace of digested mES cell DNA together with the MS signals from T and fC. The UV detection has a general delay of 0.2 min and was adjusted manually to the ion current. B) High-resolution mass data of fC. C) Fragment mass data from MS/MS and MS^3 studies proving the structure of fC. D) Fragmentation pattern of fC in the MS^2 and MS^3 experiments.

in mES cell DNA has the structure of fC. This base is present at significant levels in mES cells and in traces in mEB cell DNA.

We next quantified the amount of fC base in the mES sample (Figure 2A). To this end we co-injected a defined amount of synthetic fC together with digested mES cell DNA and integrated the ion currents of the combined fC signal. In addition, we measured the fC ion current of the added amount of synthetic fC alone and compared the values to the fC ion current measured in the mES sample. The integrals allowed us to estimate the fC level to be around 0.02%/G (Figure 2B). This is a surprisingly high value compared with the mC levels (Figure 2C) as well as with the hmC levels, which we determined by using isotope spiking to be around 0.39% hmC/G (Figure 2D).[10a,12a] Consequently, every 10th to 20th hmC base in the mES DNA is oxidized to fC, which shows that this oxidation is a significant process.

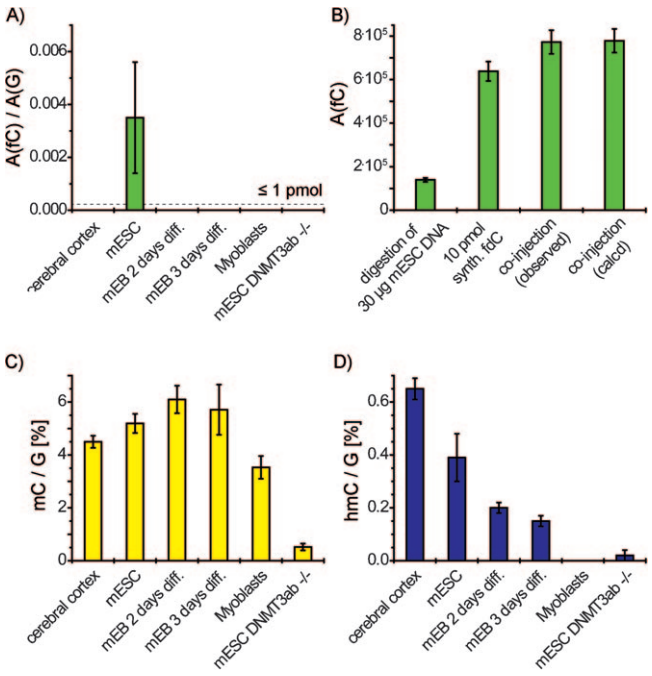


Figure 2. A) Correlation of the mass signal of fC and the UV signal of G in mES cell DNA, in DNA from nerve tissue, in DNA from mEBs after 2 and 3 days, in DNA from cultured myoblasts, and in DNA from mES lacking DNMT3a and -3b. The detection limit of fC was determined to be ≤ 1 pmol. B) Co-injection studies of fC with digested embryonic stem cell DNA indicated an amount of approx. 2 pmol, which corresponds to a level of 0.02%/G. C) Quantitative levels of mC in mES cell DNA, in DNA from nerve tissue, in DNA from mEBs after 2 and 3 days, in DNA from cultured myoblasts, and in DNA from mES lacking DNMT3a and -3b measured by quantitative mass spectrometry using an isotopically labeled mC standard.[10a,12a] D) Quantitative levels of hmC in mES cell DNA, in DNA from nerve tissue, in DNA from mEBs after 2 and 3 days, in DNA from cultured myoblasts, and in DNA from mES lacking DNMT3a and -3b measured by quantitative mass spectrometry using an isotopically labeled hmC standard.[10a,12a] Green: fC, yellow: mC, blue: hmC.

We next performed mass spectrometry experiments to study the presence of the hydrate form of fC (fC-hydrate) since further oxidation of fC to caC or elimination of a formyl group would require the addition of water to the formyl group (Scheme 1A). We indeed detected the hydrate form in DNA containing synthetic fC at a level of about 0.5% (retention time of the hydrate = 21.7 and 26.5 min; $m/z_{found} = 274.1037$, $m/z_{calcd} = 274.1034$; see the Supporting Information) which is high enough to enable either direct elimination of a formyl group (Scheme 1) or further oxidation. To investigate the presence of the further oxidized compound caC in the DNA samples from mES and mEB cells, we performed additional MS studies that were also extended to a search for the hmC- and fC-derived deamination products hmU and fU. However, signals corresponding to caC, hmU, and fU were not found. In additional MS^n experiments, fragmentation products characteristic for all these compounds were also not detected, which shows that if these compounds are present, their levels are below our detection limit.

We performed two more experiments to correlate the fC levels with the mC and hmC levels. Firstly we quantified the hmC levels in various DNA samples and secondly we studied the levels of mC, hmC, and fC in DNMT3a/3b double knock-out mES cells. Since the DNMT enzymes are needed to methylate C to mC, we hoped to learn if the newly discovered base fC is generated de novo from C (by a formylation reaction) or whether it is created from mC by stepwise oxidation via hmC. In agreement with earlier studies, we observed the highest hmC levels in nerve cell DNA (hmC/G = 0.65%). The levels of hmC in mES cell DNA are significantly lower and they decrease with differentiation (in accordance with the results of Szwagierczak et al.[14]), while the mC levels increase. A level of hmC/G = 0.39% was measured in mES cell DNA. Here also, the fC level is the highest with fC/G = 0.02%. In mEB cells, the hmC levels are hmC/G = 0.2% after 2 days of differentiation and hmC/G = 0.15% after 3 days of differentiation. The fC compound was only detected in trace amounts in both sets of mEB cells by MS^n studies. A stronger difference can be seen in the DNMT3a/3b double knock-out cells. Here, the mC levels are greatly reduced to mC/G = 0.5%, compared to 5–6% determined in the mES and mEB cells. The mC level in cultured C2C12 myoblasts is slightly lower, in agreement with other cell lines,[12a] but still at around 4%, thus showing that C methylation is, as expected, strongly hindered in the DNMT double knock-out mES cells. The same trend can be observed when studying the hmC levels. The DNMT3a/3b –/– mES cells contain practically no hmC (0.02%) and also no fC. These experiments show that fC is likely produced from mC via hmC through further oxidation. We currently speculate that the TET enzymes may convert mC into fC by iterative oxidation, as it is found for related T7H enzymes that catalyze the stepwise oxidation of thymine to 5-formyluracil.[15]

Since the formyl group of the fC compound was shown to be reactive (hydrate formation with water), we next investigated the possibility of reacting the fC base in DNA with a reagent that would allow isolation[16] of fC-containing DNA fragments for sequencing[17] and more-sensitive detection of fC. Since formyl groups react selectively with hydroxylamines to give stable oxime derivatives,[18] we treated the mES DNA with the biotin-hydroxylamine reagent (Figure 3A).[19] After

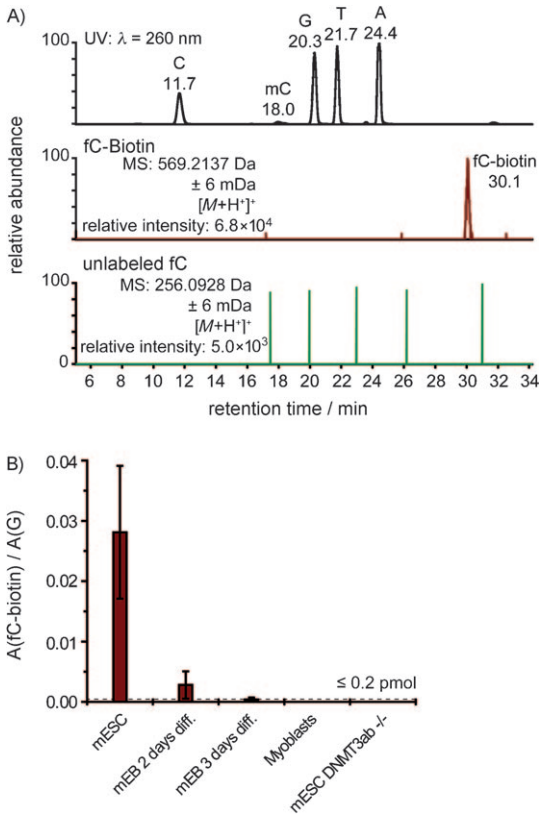


Figure 3. A) HPLC and MS signals of biotin-labeled fC obtained after treatment of embryonic stem cell DNA with the biotin hydroxylamine followed by DNA digest. The lowest trace shows only the background noise, specific signals for residual fC were not observed. B) Relative amount (mass area of fC-biotin/dG) of fC in mES cell DNA, DNA from mEBs after 2 and 3 days, DNA from cultured myoblasts, and in DNA from mES containing a double knock-out in DNMT3a/3b. The detection limit of fC-biotin using quantitative mass spectrometry is $\leq$0.2 pmol, and thus five times lower than that of unlabeled fC.

24 h of incubation (pH 5.5, 25 °C, *p*-methoxyaniline/NaOAc buffer),[20] the converted DNA was isolated and fully digested (see the Supporting Information). In a parallel experiment, we also added the biotin reagent to synthetic DNA in which the fC compound was synthetically incorporated by using a newly developed phosphoramidite building block (see the Supporting Information and for alternative synthetic strategies[21]). MS analysis of both digests showed the appearance of only a single new MS signal derived from the biotin-labeled fC nucleoside (fC-biotin). To our surprise, we no longer observed a signal for fC, thus showing that the reaction allows not only the highly selective but also complete (>90%) labeling of the fC nucleobase in genomic mES material (Figure 3A). Most importantly, the fC-biotin derivative produced a strongly increased MS signal, which allowed us to study the presence of fC in the mEB cell DNA in more detail. Indeed, after derivatization we could detect signals for the biotinylated fC base in mEB DNA (Figure 3B) directly in the mass spectrometer.

In summary, we provide here direct evidence for the presence of 5-formylcytosine (fC) in DNA isolated from mES und mEB cells. The fC levels were found to dramatically decrease with ongoing differentiation. Interestingly, we do not detect the fC compound in DNA isolated from neuronal cells, which contain the highest amounts of hmC. We explain this result on the basis of data from a recent study by Song and co-workers,[9] who showed that active demethylation in adult brain cells proceeds likely through deamination of hmC to hmU followed by removal of the hmU base by the base excision repair pathway. Thus, fC is in this respect a clear marker nucleoside for the development of mES cells. It has not escaped our notice that the oxidative demethylation of methylcytosine via 5-formylcytosine we have postulated, immediately suggests a possible globally acting epigenetic control mechanism.

Received: June 8, 2011
Published online: June 30, 2011

Keywords: active demethylation · epigenetics · 5-formylcytosine · 5-hydroxymethylcytosine · mass spectrometry

[1] a) R. Bonasio, S. J. Tu, D. Reinberg, *Science* **2010**, *330*, 612–616; b) P. A. Jones, D. Takai, *Science* **2001**, *293*, 1068–1070; c) J. A. Law, S. E. Jacobsen, *Nat. Rev. Genet.* **2010**, *11*, 204–220.

[2] a) K. Iqbal, S. G. Jin, G. P. Pfeifer, P. E. Szabo, *Proc. Natl. Acad. Sci. USA* **2011**, *108*, 3642–3647; b) M. Wossidlo, T. Nakamura, K. Lepikhov, C. J. Marques, V. Zakhartchenko, M. Boiani, J. Arand, T. Nakano, W. Reik, J. Walter, *Nat. Commun.* **2011**, *2*, 241; c) P. Hajkova, S. Erhardt, N. Lane, T. Haaf, O. El-Maarri, W. Reik, J. Walter, M. A. Surani, *Mech. Dev.* **2002**, *117*, 15–23.

[3] P. Hajkova, S. J. Jeffries, C. Lee, N. Miller, S. P. Jackson, M. A. Surani, *Science* **2010**, *329*, 78–82.

[4] a) C. A. Miller, J. D. Sweatt, *Neuron* **2007**, *53*, 857–869; b) J. J. Day, J. D. Sweatt, *Nat. Neurosci.* **2010**, *13*, 1319–1323.

[5] a) S. C. Wu, Y. Zhang, *Nat. Rev. Mol. Cell Biol.* **2010**, *11*, 607–620; b) J.-K. Zhu, *Annu. Rev. Genet.* **2009**, *43*, 143–166.

[6] 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; c) M. Münzel, D. Globisch, T. Carell, *Angew. Chem. Int. Ed.* **2011**, DOI: 10.1002/anie.201101547.

[7] a) G. Ficz, M. R. Branco, S. Seisenberger, F. Santos, F. Krueger, T. A. Hore, C. J. Marques, S. Andrews, W. Reik, *Nature* **2011**, *473*, 398–402; b) S. Ito, A. C. D'Alessio, O. V. Taranova, K. Hong, L. C. Sowers, Y. Zhang, *Nature* **2010**, *466*, 1129–1133; c) W. A. Pastor, et al., *Nature* **2011**, *473*, 394–397; d) K. Williams, J. Christensen, M. T. Pedersen, J. V. Johansen, P. A. Cloos, J. Rappsilber, K. Helin, *Nature* **2011**, *473*, 343–348; e) H. Wu, A. C. D'Alessio, S. Ito, K. Xia, Z. Wang, K. Cui, K. Zhao, Y. Eve Sun, Y. Zhang, *Nature* **2011**, *473*, 389–393; f) K. P. Koh, et al., *Cell Stem Cell* **2011**, *8*, 200–213; g) J. Walter, *Cell Stem Cell* **2011**, *8*, 121–122; h) Y. Xu, F. Wu, L. Tan, L. Kong, L. Xiong, J. Deng, A. J. Barbera, I. Zheng, H. Zhang, S. Huang, J. Min, T. Nicholson, T. Chen, G. Xu, Y. Shi, K. Zhang, Y. G. Shi, *Mol. Cell* **2011**, *42*, 451–461.

[8] a) S. H. Feng, S. E. Jacobsen, W. Reik, *Science* **2010**, *330*, 622–627; b) W. Reik, W. Dean, J. Walter, *Science* **2001**, *293*, 1089–1093.

[9] J. U. Guo, Y. Su, C. Zhong, G.-l. Ming, H. Song, *Cell* **2011**, *145*, 423–434.

[10] a) D. Globisch, M. Münzel, M. Müller, S. Michalakis, M. Wagner, S. Koch, T. Brückl, M. Biel, T. Carell, *PLoS One* **2010**, *5*, e15367; b) C. Loenarz, C. J. Schofield, *Chem. Biol.* **2009**, *16*, 580–583.

[11] T. L. Arnyes, B. M. Wood, K. Chan, J. A. Gerlt, J. P. Richard, *J. Am. Chem. Soc.* **2008**, *130*, 1574 – 1575.
[12] a) M. Münzel, D. Globisch, T. Brückl, M. Wagner, V. Welzmiller, S. Michalakis, M. Müller, M. Biel, T. Carell, *Angew. Chem.* **2010**, *122*, 5503 – 5505; *Angew. Chem. Int. Ed.* **2010**, *49*, 5375 – 5377; b) T. Le, K.-P. Kim, G. Fan, K. F. Faull, *Anal. Biochem.* **2011**, *412*, 203 – 209; c) M. Münzel, D. Globisch, C. Trindler, T. Carell, *Org. Lett.* **2010**, *12*, 5671 – 5673.
[13] H. Cao, Y. Wang, *J. Am. Soc. Mass Spectrom.* **2006**, *17*, 1335 – 1341.
[14] A. Szwagierczak, S. Bultmann, C. S. Schmidt, F. Spada, H. Leonhardt, *Nucleic Acids Res.* **2010**, *38*, e181.
[15] a) C.-K. Liu, C.-A. Hsu, M. T. Abbott, *Arch. Biochem. Biophys.* **1973**, *159*, 180 – 187; for reviews about T7H enzymes, see b) J. M. Simmons, T. A. Muller, R. P. Hausinger, *Dalton Trans.* **2008**, 5132 – 5142; c) J. M. Simmons, D. J. Koslowsky, R. P. Hausinger, *Exp. Parasitol.* **2010**, *124*, 453 – 458.
[16] C.-X. Song, K. E. Szulwach, Y. Fu, Q. Dai, C. Yi, X. Li, Y. Li, C.-H. Chen, W. Zhang, X. Jian, J. Wang, L. Zhang, T. J. Looney, B. Zhang, L. A. Godley, L. M. Hicks, B. T. Lahn, P. Jin, C. He, *Nat. Biotechnol.* **2011**, *29*, 68 – 72.
[17] S.-G. Jin, S. Kadam, G. P. Pfeifer, *Nucleic Acids Res.* **2010**, *38*, e125.
[18] a) L. M. Hough, F. L. Oswald, *Annu. Rev. Psychol.* **2000**, *51*, 631 – 664; b) V. Raindlová, R. Pohl, M. Sanda, M. Hocek, *Angew. Chem.* **2010**, *122*, 1082 – 1084; *Angew. Chem. Int. Ed.* **2010**, *49*, 1064 – 1066.
[19] a) J. Nakamura, V. E. Walker, P. B. Upton, S.-Y. Chiang, Y. W. Kow, J. A. Swenberg, *Cancer Res.* **1998**, *58*, 222 – 225; b) K. Kubo, H. Ide, S. S. Wallace, Y. W. Kow, *Biochemistry* **1992**, *31*, 3703 – 3708.
[20] A. Dirksen, T. M. Hackeng, P. E. Dawson, *Angew. Chem.* **2006**, *118*, 7743 – 7746; *Angew. Chem. Int. Ed.* **2006**, *45*, 7581 – 7584.
[21] a) N. Karino, Y. Ueno, A. Matsuda, *Nucleic Acids Res.* **2001**, *29*, 2456 – 2463; b) Q. Dai, C. He, *Org. Lett.* **2011**, DOI: 10.1021/ol201189n.

