5-Hydroxymethylcytosine

DOI: 10.1002/anie.201204768

Sequencing the Sixth Base (5-Hydroxymethylcytosine): Selective DNA Oxidation Enables Base-Pair Resolution**

*Peter Schüler and Aubry K. Miller**

5-hydroxymethylcytosine · DNA · enzymes · oxidation · sequencing

The methylation of cytosine (C) to give 5-methylcytosine (5mC) in mammalian DNA is an important epigenetic modification impacting development and gene expression, and has been studied for decades (Figure 1). In 2009, two groups simultaneously reported the discovery of 5-hydroxymethylcytosine (5hmC) in mammalian DNA and showed that

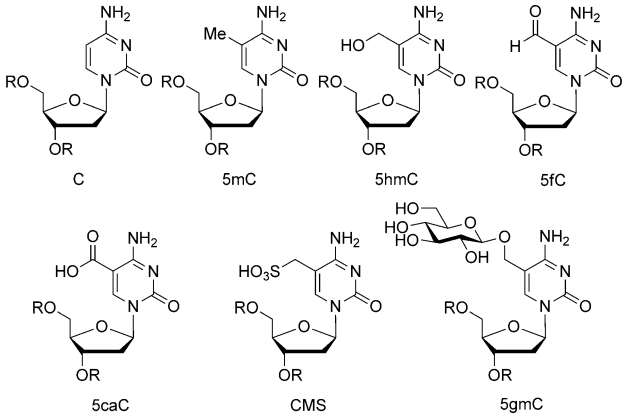
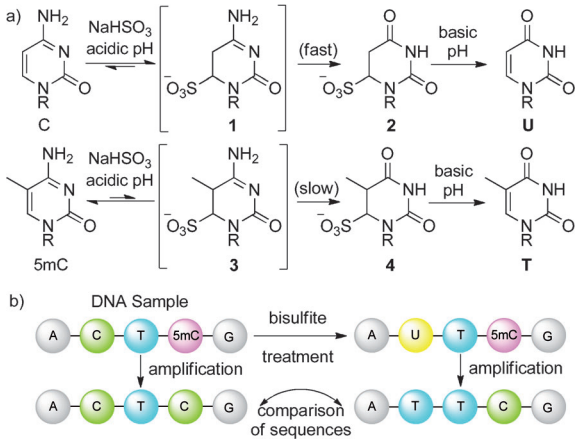


Figure 1. Natural (C, 5mC, 5hmC, 5fC, and 5caC) and unnatural (CMS and 5gmC) eukaryotic cytosine congeners. R = DNA.

the family of ten-eleven-translocation (TET) oxygenases converts 5mC to 5hmC.[1,2] Last year, the two more highly oxidized congeners of 5hmC, 5-formylcytosine and 5-carboxycytosine (5fC and 5caC respectively), were also discovered as 5mC TET oxidation products.[3] The race to disentangle the roles that these "new" DNA bases play, both as epigenetic markers and as intermediates in demethylation pathways, is on.

The bisulfite-mediated deamination of cytosines to uridines has played a crucial role in understanding DNA methylation. Shapiro and Hayatsu independently reported and subsequently performed detailed investigations of this reaction over 40 years ago on single nucleotides (Scheme 1 a).[4] They showed that bisulfite readily adds to C



Scheme 1. a) Bisulfite-mediated deamination of C and 5mC. b) PCR amplification of a DNA sample converts both C and 5mC to C. Bisulfite treatment of the same sample converts C to U while 5mC (and all other bases) remains unchanged. After amplification and sequencing, the positions that are read as C indicate where a 5mC resides. Comparison of this data to the results of a standard sequencing run then reveals the positions of C residues.

to give sulfonate **1** which then undergoes hydrolysis to **2**. Basification then promotes elimination of bisulfite to reveal the "deaminated" uridine (U) product. It was subsequently discovered that deamination of 5mC to thymidine (T), via **3** and **4**, is nearly two orders of magnitude slower than for C. This rate difference is taken advantage of in what is known as bisulfite sequencing (BS-Seq). Today, BS-Seq kits can be purchased to selectively deaminate C in the presence of 5mC on genomic DNA with high fidelity (Scheme 1 b). Comparison of normal and bisulfite sequencing data reveals the location of 5mC in DNA, and has become a powerful and now routine tool for the epigenetic community that can even be used to map entire methylomes.[5]

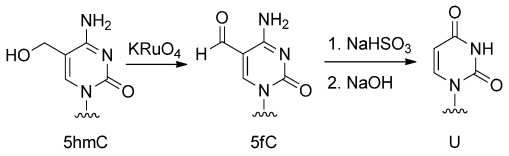
[*] Dr. P. Schüler, Dr. A. K. Miller
Cancer Drug Development
Deutsches Krebsforschungszentrum (DKFZ)
Im Neuenheimer Feld 280, 69120 Heidelberg (Germany)
E-mail: aubry.miller@dkfz.de
Homepage: http://www.dkfz.de/en/drugs/index.php

[**] We thank the Helmholtz Drug Research Initiative for funding (P.S.) and the program "epigenetics@dkfz".

After the discovery of 5hmC, it was found that BS-Seq alone is unable to distinguish between 5mC and 5hmC.[6] As originally shown by Hayatsu, 5hmC reacts with bisulfite to give cytosine 5-methylenesulfonate (CMS; Figure 1), a compound that undergoes deamination even more slowly than 5mC and is read as a C when amplified and sequenced (Scheme 2b).[7] This indicates that current genome-wide bisulfite sequencing maps are not entirely accurate and can only be corrected with the development of new, more powerful sequencing protocols. Recently, the groups of Balasubramanian[8] and He[9] reported modified BS-Seq protocols that provide base-pair resolution of 5hmC. Both of these methods rely on selective chemical transformations on genomic DNA followed by BS-Seq. In this Highlight we would like to focus on the chemistry behind these techniques.

Last year, the He group, in collaboration with Guo-Liang Xu's group, reported that 5caC behaves like C in bisulfite sequencing, meaning that it is read as T.[3b] The He group has now shown that 5mC can be oxidized all the way to 5caC in the presence of excess recombinant Tet1, while unmodified C does not react.[9] They reasoned that if they could selectively convert 5mC to 5caC in the presence of 5hmC, BS-Seq of the resulting DNA strand would identify the 5hmC loci. What was then required was a "protecting group" for 5hmC. The final protocol, which they have named Tet-assisted bisulfite sequencing (TAB-Seq), relies on two enzymatic transformations (Scheme 2c). First, using β-glucosyl transferase (βGT), each 5hmC base on a strand of genomic DNA is protected as β-glucosyl-5-hydroxymethylcytosine (5gmC; Figure 1). Second, the DNA is treated with excess Tet1 to oxidize 5mC loci to 5caC. Subsequent bisulfite treatment converts all C and presumably all 5caC bases (vide infra) to U while the 5gmC bases remain unaffected. After amplification (5gmC amplifies to C) and sequencing, the positions that are read as C indicate where a 5hmC resides. Comparison of this data to the results of a standard BS-Seq run then reveals the positions of 5mC residues.

Balasubramanian's approach is conceptually similar to He's method in that selective oxidation of a specific cytosine congener is followed by BS-Seq. The method hinges on the Balasubramanian group's finding that 5hmC can be oxidized with $KRuO_4$ to 5fC, a compound that undergoes bisulfite-mediated deformylative deamination to yield U (Scheme 3). The final technique, referred to as oxidative bisulfite sequencing (oxBS-Seq), begins with $KRuO_4$ oxidation of all 5hmC residues on a strand of DNA to 5fC (Scheme 2a). Remarkably, this chemical oxidation is selective for 5hmC, even on genomic DNA. Subsequent bisulfite treatment converts all C loci, and all of the newly formed 5fC residues, to U. After amplification and sequencing, the positions that are read as C indicate where a 5mC unit resides. Comparison of this data to the results of a standard BS-Seq run then reveals the positions of 5hmC residues.

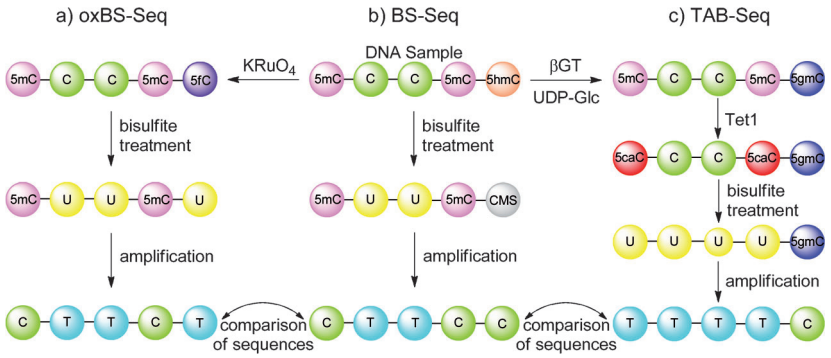


Scheme 3. Oxidation of 5hmC followed by bisulfite-mediated deformylative deamination of 5fC.

While oxBS-Seq (oxidation of 5hmC) and TAB-Seq (protection of 5hmC followed by exhaustive 5mC oxidation) generate different outputs, they ultimately yield the exact same information (Figure 2). Therefore, the ease and reliability of the two techniques will ultimately determine which protocol will predominate. The oxBS-Seq protocol is operationally simpler: no "protecting groups", only one chemical transformation, and one "purification" are required before bisulfite treatment. The sole reagent, $KRuO_4$, is a standard chemical and easy to handle on the small scale required. The TAB-Seq protocol, on the other hand, relies on two enzymatic

Base	Standard Sequence	Bisulfite Sequence	TAB Sequence	oxBS Sequence
C	C	T	T	T
5mC	C	C	T	C
5hmC	C	C	C	T

Figure 2. Comparison of either TAB-Seq or oxBS-Seq with standard and bisulfite sequencing data logically determines the location of C, 5mC, and 5hmC.



Scheme 2. Comparison of the a) oxBS-Seq, b) BS-Seq, and c) TAB-Seq protocols.

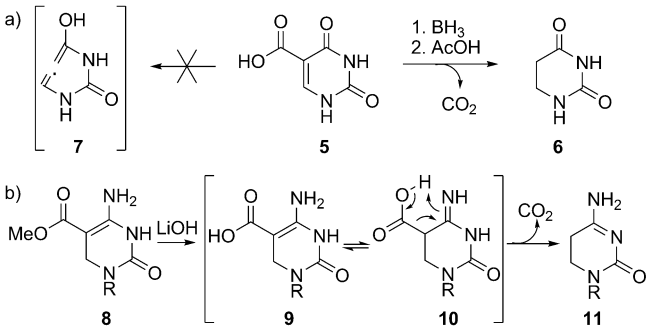
transformations, each of which requires a purification step before bisulfite treatment. While the production and use of βGT is well established, the same cannot be said about Tet1, yet. In their study, the He group relied on Tet1 expression in insect cells and it is unclear whether enzyme production can be streamlined to consistently produce a high-quality tool at a competitive price.

By qPCR of a sequencing library prepared from genomic DNA, the Balasubramanian group estimated that only 0.5% of the DNA fragments remained intact after oxBS-Seq (95% cleaved during oxidation and another 90% during bisulfite treatment). While no information on DNA degradation from the TAB-Seq procedure was given, it seems unlikely that βGT or Tet1 would significantly contribute to DNA degradation. As both techniques can be performed with 2 μg of genomic DNA, however, the high level of DNA degradation in oxBS-Seq may actually be negligible.

To measure the efficiency of their methods, both groups used quantitative sequencing techniques. The Balasubramanian group reported that roughly 5% of 5hmC residues are misread as 5mC with oxBS-Seq, while the He group determined a value of approximately 8% for TAB-Seq. The He group also found a reduced glucosylation rate in the case of closely associated 5hmC positions, presumably due to steric effects. They found that a distance of three bases between neighboring 5hmCs is enough to maintain a protection ratio similar to that of solitary 5hmCs. In the extreme case of a single G separating two 5hmCs, βGT performed at 90% of the level of solitary 5hmCs. While one might speculate that sterics would play a lesser role in oxBS-Seq, this type of benchmarking analysis would also be illuminating for this technique.

We expect improvement in the overall efficiencies of these methods in the future, but both already offer a deeper insight into methylation/hydroxymethylation patterns. Although oxBS-Seq constitutes a rather harsh method, the fact that all reagents are commercially available gives it an edge over TAB-Seq, for which enzyme production has to be mastered beforehand.

While not directly addressed in the He manuscript, it is extremely likely that bisulfite treatment of 5caC-containing DNA results in decarboxylation and deamination to give U, a reaction demonstrated by Isono on 5-carboxycytosine base as early as 1972.[10] A fascinating feature of both new sequencing protocols therefore emerges: the bisulfite-mediated removal of one carbon from a cytosine derivative under mild conditions either as a decarboxylation (TAB-Seq) or a deformylation (oxBS-Seq). The mechanisms by which these reactions operate have not yet been completely elucidated, although works by Pal and Carell shed some light on the decarboxylation of 5caC. In 1984, Pal showed that treatment of 5-carboxyuracil (**5**) with borane provides 5,6-dihydrouracil (**6**) (Scheme 4a).[11] Presumably, saturation (via hydroboration) of the C5–C6 double bond in **5** gives an intermediate that, resembling a β-ketoacid, loses CO_2 spontaneously. The presence of the C5–C6 double bond in **5** prohibits spontaneous decarboxylation through a similar pathway because the hypothetical first intermediate (**7**) would incorporate an sp-hybridized carbon in a six-membered ring.



Scheme 4. a) Saturation of the C5–C6 double bond in **5** results in decarboxylation. b) Saponification of **9** leads to decarboxylation. R = 3′,5′-bis(OTBS)-2′-deoxyribose, TBS = *t*-butyldimethylsilyl.

Recently, Carell showed that similar chemistry occurs with 5-carboxycytosine derivatives.[12] Saponification of the methyl ester of dihydro-5caC **8** at room temperature gave **11** in 10% yield, presumably through decarboxylation of tautomer **10** (Scheme 4b). The Carell group then demonstrated that a small percentage of 5caC loci in DNA spontaneously decarboxylate in the presence of cysteine, an amino acid that could be capable of temporarily saturating the C5–C6 double bond.

Bisulfite very effectively saturates the C5–C6 double bond of cytosines and may promote the decarboxylation of 5caC in a similar manner; however, to the best of our knowledge, mechanistic details for the bisufite-mediated decarboxylation of 5caC and deformylation of 5fC are sparse and require investigation. Stereochemical issues could play a role as bisulfite is known to add with complete stereospecificity across the C5–C6 double bond of pyrimidine bases, albeit with poor diastereoselectivity with respect to the chiral ribose backbone. This suggests, particularly in the case of 5fC, that many diastereomeric intermediates are formed during the course of the reactions. Regardless of the mechanisms operating, the selective oxidation of DNA coupled to the surprisingly powerful chemistry of sodium bisulfite has enabled, for the first time, sequencing of 5hmC with single-base-pair resolution. The epigenetic research community will certainly benefit from these breakthroughs. We expect to see continued improvements in the future and perhaps even sequencing methods to identify the most scarce cytosine congeners, 5fC and 5caC.

Received: June 18, 2012
Published online: September 26, 2012

[1] S. Kriaucionis, N. Heintz, *Science* **2009**, *324*, 929–930; 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.

[2] 5hmC was first reported to exist in mammalian DNA in 1972: N. W. Penn, R. Suwalski, C. O'Riley, K. Bojanowski, R. Yura, *Biochem. J.* **1972**, *126*, 781–790. As this finding could not be reproduced, the topic remained largely dormant for decades.

[3] 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; c) T. Pfaffeneder, B. Hackner, M. Truss, M. Muenzel, M. Mueller, C. A. Deiml, C. Hagemeier, T. Carell, *Angew. Chem.* **2011**, *123*, 7146 – 7150; *Angew. Chem. Int. Ed.* **2011**, *50*, 7008 – 7012.

[4] R. Shapiro, R. E. Servis, M. Welcher, *J. Am. Chem. Soc.* **1970**, *92*, 422 – 424; H. Hayatsu, Y. Wataya, K. Kai, *J. Am. Chem. Soc.* **1970**, *92*, 724 – 726.

[5] R. Lister, M. Pelizzola, R. H. Dowen, R. D. Hawkins, G. Hon, J. Tonti-Filippini, J. R. Nery, L. Lee, Z. Ye, Q.-M. Ngo, L. Edsall, J. Antosiewicz-Bourget, R. Stewart, V. Ruotti, A. H. Millar, J. A. Thomson, B. Ren, J. R. Ecker, *Nature* **2009**, *462*, 315 – 322; S. Balasubramanian, *Angew. Chem.* **2011**, *123*, 12612 – 12616; *Angew. Chem. Int. Ed.* **2011**, *50*, 12406 – 12410.

[6] Y. Huang, W. A. Pastor, Y. Shen, M. Tahiliani, D. R. Liu, A. Rao, *PLoS One* **2010**, *5*, e88888; S.-G. Jin, S. Kadam, G. P. Pfeifer, *Nucleic Acids Res.* **2010**, *38*, e125.

[7] H. Hayatsu, M. Shiragami, *Biochemistry* **1979**, *18*, 632 – 637.

[8] M. J. Booth, M. R. Branco, G. Ficz, D. Oxley, F. Krueger, W. Reik, S. Balasubramanian, *Science* **2012**, *336*, 934 – 937.

[9] M. Yu, G. C. Hon, K. 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.

[10] K. Isono, S. Suzuki, M. Tanaka, T. Nanbata, K. Shibuya, *Agric. Biol. Chem.* **1972**, *36*, 1571 – 1579.

[11] C. Ghosh, D. G. Schmidt, B. C. Pal, *J. Org. Chem.* **1984**, *49*, 5256 – 5257.

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