Identification of sequence specificity of 5-methylcytosine oxidation by Tet1 protein with high-throughput sequencing

Seiichiro Kizaki,[a] Anandhakumar Chandranan,[a] and Hiroshi Sugiyama*[a,b]

Abstract: Tet (ten-eleven translocation) family proteins have the ability to oxidize 5-methylcytosine (mC) to 5-hydroxymethylcytosine (hmC), 5-formylcytosine (fC), and 5-carboxycytosine (caC). However, the oxidation reaction property of Tet protein is not understood completely. Evaluation of genomic-level epigenetic changes by Tet protein requires unbiased identification of the highly selective oxidation sites. In this study, we used high-throughput sequencing to investigate the sequence specificity of mC oxidation by Tet1 protein. A 6.6×10^4-member mC-containing randomized DNA sequence library was constructed. The library was subjected to Tet-reactive-specific pulldown followed by high-throughput sequencing. Analysis of the obtained sequencing data identified the Tet1-reactive sequences. We identified mCpG as a highly reactive sequence of Tet1. The identified reactive site correlated well with the regions containing hmC in the mESC genome.

Introduction

Methylation and demethylation of cytosine in DNA has attracted attention recently as epigenetic control of gene expression. Since the discovery of the ability of Tet (ten-eleven translocation) family proteins to convert 5-methylcytosine (mC) to 5-hydroxymethylcytosine (hmC) and further to 5-formylcytosine (fC) and 5-carboxy cytosine (caC) (Fig. 1).[1–3] In fact, many methods to detect hmC, fC, and caC in genomic DNA have been developed,[4–11] and all of these oxidized derivatives of mC are known to exist in mammalian tissues.[1,2,12–16] Thus, mechanism of demethylation has drawn current interest because of their essential roles in epigenetic gene regulation, embryogenesis and cellular reprograming.[17–19]

Wu *et al.* performed a ChIP-seq study using antibody specific for Tet1 and revealed that Tet1 binds preferentially to CpG-rich sequences in mouse ES cells.[20] Although this study was very informative about many aspects, the process examined was limited to the binding site of Tet1. We believe that it is also important to reveal the Tet-reactive sites to identify the reversible methylation region, and we developed a method to reveal these sites.

[a] S. Kizaki, A. Chandran, Prof. Dr. H. Sugiyama
Department of Chemistry, Kyoto University
Kitashirakawa-Oiwake-cho, Sakyo-ku, Kyoto-shi, Kyoto, 606-8502 (Japan)
E-mail: hs@kuchem.kyoto-u.ac.jp

[b] Prof. Dr. H. Sugiyama
Institute for Integrated Cell-Material Sciences (iCeMS), Kyoto University
Yoshida-ushinomiyacho, Sakyo-ku, Kyoto-shi, Kyoto, 606-8501 (Japan)

Supporting information for this article is given via a link at the end of the document.

The advent of high-throughput sequencing has initiated great progress in the study of genomes. The application of high-throughput sequencing is not limited to the analysis of genomic information. For example, Meier *et al.* used high-throughput sequencing to reveal the binding sequence of pyrrole–imidazole polyamide, which is known to bind to the minor groove of DNA in a sequence-specific manner.[21] Anandhakumar *et al.* reviewed the importance of high-throughput sequencing in the chemical biology.[22]

In this study, we used high-throughput sequencing to identify the specific DNA sequence with which Tet1 prefers to react.

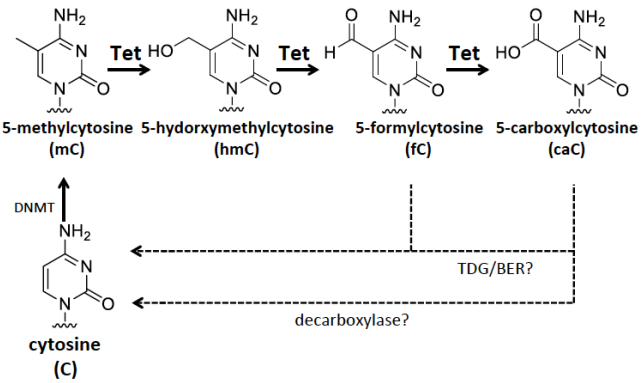


Figure. 1 Cytosine, 5-methylcytosine, and oxidation products of 5-methylcytosine by Tet protein. Cytosine is methylated by DNA methyltransferase (DNMT), and 5-methylcytosine is oxidized by Tet protein to form 5-hydroxymethylcytosine (hmC), 5-formylcytosine (fC), and 5-carboxycytosine (caC). TDG: thymine DNA glycosylase, BER: base excision repair

Results and discussion

To identify the Tet1-reactive site, we first prepared a pool of DNAs containing mC flanked by random 8-mer sequences (Fig. 2). We incubated this pool of DNAs with three different concentrations of Tet1 protein. Only Tet1-reacted DNAs were enriched with anti-hmC monoclonal antibody, because mC to hmC conversion occurs efficiently only in the regions where Tet1 protein reacts easily. The enriched DNAs were sequenced further by high-throughput sequencing.

The binding efficiency of the anti-hmC monoclonal antibody toward the oligomers was assessed using 6-mer DNA containing either mC or hmC. The results are shown in Supplementary Figure 1.[23] The results clearly indicated the specificity of antibody toward hmC even with short oligomers.

The sequenced reads were processed by splitting based on barcodes and filtering the low-quality reads. The enrichment sequence and its corresponding common nucleotide arrangement pattern (motif) were derived with respect to the control using Bind-n-Seq analysis and the MEME algorithm (Table 1, Supplementary Tables 1 and 2).[24–26] To perform *de*

novo motif analysis with more stringent cutoffs to minimize background, we defined the highly enriched sequences as intermediate motifs (Supplementary Figure 2).[24] Each motif produced different sequences, but the sixth position (next to hmC toward the 3′ end) remained with highly similar bases for the pool of DNA treated with a low concentration (0.0663 μM) of Tet1 protein. The reads corresponding to intermediate motifs were merged together and used to define the final motif (Fig. 3).

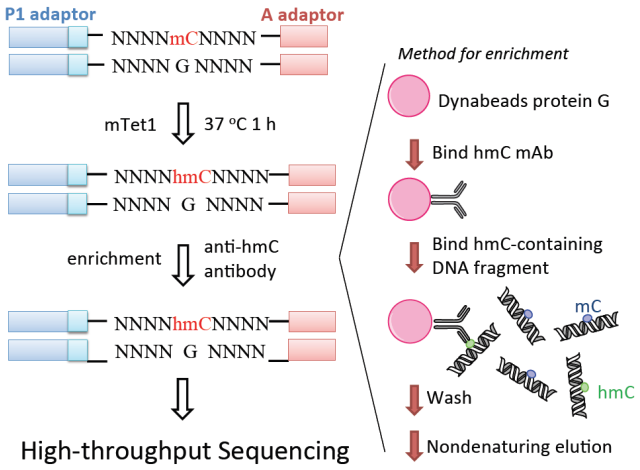


Figure 2. Treatment of the DNA library with Tet1 protein, followed by enrichment with anti-hmC antibody and high-throughput sequencing. (mAb: monoclonal antibody)

When the DNA pool was incubated with a high (6.63 μM) or medium (0.663 μM) concentration of Tet1 protein, almost no sequence specificity was observed for the mC-oxidation reaction by Tet1 protein (Fig. 3(a) and (b)). However, when the DNA pool was incubated with a low (0.0663 μM) concentration of Tet1 protein, obvious sequence specificity for the 3′-region of hmC was observed (Fig. 3(c)). The 3′-region contained G:C in a 3:1 ratio. These results suggests that, at higher concentrations, Tet1 protein oxidize mC regardless of the flanking DNA sequence, but at lower concentrations, Tet1 protein reads the flanking DNA sequence for mC oxidation. In eukaryotic cells, it is possible that the local concentration of Tet1 protein is strictly regulated by various kinds of Tet1-associating proteins to control the oxidation reaction of mC.[27,28]

To assess the Tet1-reactive site at the genome scale, we obtained data for hmC-specific SMRT sequencing of the mESC genome.[29] Using these data, we identified motifs specific to hmC and to hemi-hmC in both the + and – DNA strands; the motifs are shown in Supplementary Figure 3. The results correlated well with the identified reactive sites: the nonselective 5′-region and the selective 3′-region of mC.

From our study, we conclude that the mCpG sequence can be a good reactive substrate for Tet1 protein. This result also suggests that Tet1 protein can reverse the methylation at mCpG regions. To confirm this conversion at the genomic scale, we plan to perform genomic-level Tet1-reactive-specific pull-down and sequencing assays.

Table 1. Relative enrichment of 9-base pair "k"mers. (only top 10 sequences are shown). Randomized DNAs were incubated with 0.0663 μM Tet protein. The C in the center is mC in the original DNA sequence and hmC in the enriched sequence.

Rank	Sequence	Enrichment
1	ATGG**C**GTTT	44.801
2	ATGG**C**GTTG	43.013
3	TCAT**C**GCTT	42.119
4	TCTA**C**GCTT	40.331
5	ATGG**C**GGTT	40.331
6	TCCT**C**CCTT	37.650
7	GCTC**C**GTCC	36.756
8	TCTG**C**CCTT	36.756
9	TCCG**C**GTTC	35.862
10	TCCA**C**CCGT	35.862

Conclusions

We used high-throughput sequencing to identify the reactive sequence of Tet1 protein. This is the first study to use high-throughput sequencing to reveal the sequence specificity of this enzymatic reaction. This novel method is applicable for identifying the reactive sequences of other enzymes that act on DNA. This method with a standardized reactive enzyme concentration can be used in studies at the genomic and cellular levels.

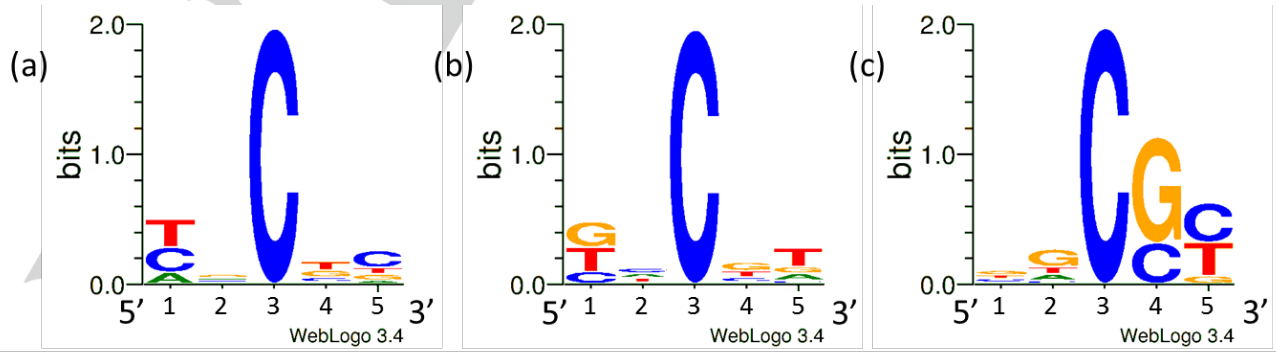


Figure 3. Reactive motif (final motif) identified from the intermediate motifs derived using Bind-n-Seq through next-generation sequencing. Randomized DNAs were incubated with (a) 6.63 μM, (b) 0.633 μM, or (c) 0.0633 μM Tet protein. The C in the center is mC in the original DNA sequence and hmC in the enriched sequence.

Experimental Section

To identify Tet reactive site we used broad context of randomized DNA sequences with mC. For the DNA sequence library construction, previously reported Bind-n-Seq experiments [21,22,30] were customized with the modifications suitable for Tet reaction and Ion semiconductor sequencer. The experimental strategy consists of three major phases, as follows

1) mTet1 and its oxidative substrate DNA preparation:

mTet1 active domain (1367-2039) was purchased from Wisegene (USA), stocked in 20 mM HEPES (pH 7.4), NaCl 50 mM, glycerol 50%.

Ion semiconductor sequencer specific adapters ligated 9 mer randomized DNA library was synthesized. The ssDNA library contains 8 randomized (N) nucleotides and one mC in the 5th position (Library sequence details provided in the supplementary information). Duplex Tet reactive DNA library was obtained through primer extension of ssDNA library with adapter specific PE primer in 200 µL reaction containing 1xGoTaq Green Master Mix (Promega) with 500 µM $MgCl_2$. Reactions were performed 95°C /2 min, 60°C /1 min, 70°C /5 min and then cooled down to 4 °C using Bio-Rad T100 thermal cycler.

2) mTet1 oxidation specific pulldown:

To prepare hmC antibody-supported magnetic beads, 50 µL of Dynabeads Protein G (Life Technologies, USA) was transferred to 0.2 mL PCR tube, and stock solution was removed after magnetic separation. Magnetic beads was dissolved in 12 µL of monoclonal 5-hydroxymethylcytosine antibody (Active Motif, USA) and 188 µL of PBS (pH 7.4) with 0.02% Tween-20, and then incubated at 4 °C for 10 min with rotation. To check whether hmC-containing DNA fragment is selectively attached to prepared hmC antibody-supported magnetic beads, 1 µL of 200 µM 5'-d(CGmCGCG)-3', 1 µL of 200 µM 5'-d(CGhmCGCG)-3', 48 µL of Milli-Q, and 200 µL of IP buffer (0.01% SDS, 1.1% Triton X-100, 1.2 mM EDTA, 16.7 mM Tris–HCl (pH 8.0), 167 mM NaCl) was added, then incubated at 4 °C for 18 hours. After incubation, the sample tube was placed on a magnetic stand. Then the supernatant was taken and analyzed by a high-performance liquid chromatography (HPLC) system (JASCO) equipped with a reversed-phase ODS column CHEMCOBOND 5-ODS-H (Chemco Scientific). Elution was done with 50 mM ammonium formate containing 0–3% acetonitrile in a linear gradient at a flow rate of 1.0 mL min^{-1} for 30 minutes, at 40 °C.

mC-containing randomized duplexed DNAs (1.08 µM) were incubated with 0.0663, 0.663, or 6.63 µM of mTet1 protein in 50 mM HEPES (pH 8.0), 100 mM NaCl, 2 mM L-ascorbic acid, 1 mM 2-oxoglutarate disodium salt hydrate, 105 µM $Fe(NH_4)_2(SO_4)_2$ $6H_2O$, 1.2 mM ATP and 2.5 mM DTT at 37 °C for 1 hour in 50 µL reaction, and then 200 µL of IP-buffer was added to the reaction.

Whole 250 µL of Tet reaction and 200 µL of hmC antibody-supported magnetic beads were mixed together and incubated with rotation at 4 °C for 24 hours. After incubation, the reaction sample was washed with 200 µL of PBS (pH 7.4) three times. The duplexed DNAs containing hmC were finally eluted with 50 mM glycine (pH 2.8) followed by neutralization with 1 M TrisHCl (pH 7.5).

3) Sequencing and data analysis:

Antibody-enriched DNA was diluted 1:10 and amplified with sequencing library Ion sequencing library amplification kit in-order to attain sufficient library for sequencing template preparation. After amplification the libraries were purified. Then the libraries were quantified with Agilent DNA High sensitivity BioAnalyzer kit (Agilent technologies, USA). Followed by the libraries were subjected to template preparation (Ion PGM^{TM} template OT2 200 kit) in Ion one touch2 system. Ion one touch ES was used for library enrichment. Ion PGM sequencer was used to sequence the enriched libraries with Ion PGM^{TM} sequencing 200 kit v2 and 318C chip (Life Technologies, USA). The sequenced reads were separated into individual files based on the unique 10-nt ion Xpress-barcode used in the library construction using the Ion torrent suit 4.2.1. High quality reads (containing only A, C, T, or G) without homopolymer error and unique random region were filtered.

The reads containing "C" at the fifth position were extracted for the further analysis. To identify the mTet1 reactive site from the enriched unique DNA sequences, mTet1 enriched data were normalised with negative control (No mTet1). The differentially enriched sequence motifs (intermediate motifs) were identified using Bind-n-Seq analysis (http://korflab.ucdavis.edu/Data sets/BindNSeq.27).[24] Based on the intermediate motif sequences the final motif was obtained. Highly enriched (mTet1 oxidative site) final motifs were created using WebLogo (http://weblogo.berkeley.edu) and further confirmed with MEME motif analysis algorithm.[24,25,31]

Acknowledgements

This work was supported by JSPS KAKENHI (Grant Number 24225005), "Basic Science and Platform Technology Program for Innovative Biological Medicine" and "JSPS-NSF International Collaborations in Chemistry (ICC)" to HS.

Keywords: Tet • 5-hydroxymethylcytosine • 5-methylcytosine • high-throughput sequencing

References

[1] M. Tahiliani, K. P. Koh, Y. Shen, W. a Pastor, H. Bandukwala, Y. Brudno, S. Agarwal, L. M. Iyer, D. R. Liu, L. Aravind, et al., *Science* **2009**, *324*, 930–5.

[2] S. Ito, L. Shen, Q. Dai, S. C. Wu, L. B. Collins, J. a Swenberg, C. He, Y. Zhang, *Science* **2011**, *333*, 1300–3.

[3] Y.-F. He, B.-Z. Li, Z. Li, P. Liu, Y. Wang, Q. Tang, J. Ding, Y. Jia, Z. Chen, L. Li, et al., *Science* **2011**, *333*, 1303–7.

[4] M. Yu, G. C. Hon, K. E. Szulwach, C.-X. Song, L. Zhang, A. Kim, X. Li, Q. Dai, Y. Shen, B. Park, et al., *Cell* **2012**, *149*, 1368–80.

[5] M. Yu, G. C. Hon, K. E. Szulwach, C.-X. Song, P. Jin, B. Ren, C. He, *Nat. Protoc.* **2012**, *7*, 2159–2170.

[6] L. Zhang, K. E. Szulwach, G. C. Hon, C.-X. Song, B. Park, M. Yu, X. Lu, Q. Dai, X. Wang, C. R. Street, et al., *Nat. Commun.* **2013**, *4*, 1517.

[7] M. J. Booth, M. R. Branco, G. Ficz, D. Oxley, F. Krueger, W. Reik, S. Balasubramanian, *Science* **2012**, *336*, 934–7.

[8] M. J. Booth, T. W. B. Ost, D. Beraldi, N. M. Bell, M. R. Branco, W. Reik, S. Balasubramanian, *Nat. Protoc.* **2013**, *8*, 1841–51.

[9] P. Schüler, A. K. Miller, *Angew. Chem. Int. Ed. Engl.* **2012**, *51*, 10704–7.

[10] C.-X. Song, K. E. Szulwach, Q. Dai, Y. Fu, S.-Q. Mao, L. Lin, C. Street, Y. Li, M. Poidevin, H. Wu, et al., *Cell* **2013**, *153*, 1–14.

[11] X. Lu, C. Song, K. Szulwach, Z. Wang, P. Weidenbacher, P. Jin, C. He, *J. Am. Chem. Soc.* **2013**, *135*, 9315–9317.

[12] S. Kriaucionis, N. Heintz, *Science* **2009**, *324*, 929–30.

[13] M. Münzel, D. Globisch, T. Brückl, M. Wagner, V. Welzmiller, S. Michalakis, M. Müller, M. Biel, T. Carell, *Angew. Chem. Int. Ed. Engl.* **2010**, *49*, 5375–7.

[14] T. Pfaffeneder, B. Hackner, M. Truß, M. Münzel, M. Müller, C. a. Deiml, C. Hagemeier, T. Carell, *Angew. Chemie* **2011**, *123*, 7146–7150.

[15] D. Globisch, M. Münzel, M. Müller, S. Michalakis, M. Wagner, S. Koch, T. Brückl, M. Biel, T. Carell, *PLoS One* **2010**, *5*, e15367.

[16] M. Münzel, D. Globisch, T. Carell, *Angew. Chem. Int. Ed. Engl.* **2011**, *50*, 6460–8.

[17] J. a Hackett, R. Sengupta, J. J. Zylicz, K. Murakami, C. Lee, T. a Down, M. A. Surani, *Science* **2013**, *339*, 448–52.

[18] J. U. Guo, Y. Su, C. Zhong, G. Ming, H. Song, *Cell* **2011**, *145*, 423–34.

[19] T. Nakamura, Y.-J. Liu, H. Nakashima, H. Umehara, K. Inoue, S. Matoba, M. Tachibana, A. Ogura, Y. Shinkai, T. Nakano, *Nature* **2012**, *486*, 415–9.

[20] H. Wu, A. C. D'Alessio, S. Ito, K. Xia, Z. Wang, K. Cui, K. Zhao, Y. E. Sun, Y. Zhang, *Nature* **2011**, *473*, 389–93.

[21] J. L. Meier, A. S. Yu, I. Korf, D. J. Segal, P. B. Dervan, *J. Am. Chem. Soc.* **2012**, *134*, 17814–22.

[22] C. Anandhakumar, S. Kizaki, T. Bando, G. N. Pandian, H. Sugiyama, *ChemBioChem* **2015**, *16*, 20–38.

[23] S. Kizaki, H. Sugiyama, *Org. Biomol. Chem.* **2014**, *12*, 104–7.

[24] A. Zykovich, I. Korf, D. J. Segal, *Nucleic Acids Res.* **2009**, *37*, e151.

[25] T. L. Bailey, M. Boden, F. A. Buske, M. Frith, C. E. Grant, L. Clementi, J. Ren, W. W. Li, W. S. Noble, *Nucleic Acids Res.* **2009**, *37*, W202–8.

[26] G. E. Crooks, G. Hon, J. M. Chandonia, S. E. Brenner, *Genome Res.* **2004**, *14*, 1188–1190.

[27] P.-F. Cartron, A. Nadaradjane, F. Lepape, L. Lalier, B. Gardie, F. M. Vallette, *Genes Cancer* **2013**, *4*, 235–41.

[28] F. T. Shi, H. Kim, W. Lu, Q. He, D. Liu, M. A. Goodell, M. Wan, Z. Songyang, *J. Biol. Chem.* **2013**, *288*, 20776–20784.

[29] C.-X. Song, T. a Clark, X.-Y. Lu, A. Kislyuk, Q. Dai, S. W. Turner, C. He, J. Korlach, *Nat. Methods* **2012**, *9*, 75–7.

[30] J. S. Kang, J. L. Meier, P. B. Dervan, *J. Am. Chem. Soc.* **2014**, *136*, 3687–94.

[31] T. D. Schneider, R. M. Stephens, *Nucleic Acids Res.* **1990**, *18*, 6097–6100.

Entry for the Table of Contents (Please choose one layout)

Layout 1:

COMMUNICATION

Text for Table of Contents

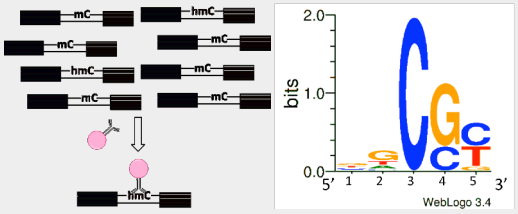
*Author(s), Corresponding Author(s)**

Page No. – Page No.

Title

Layout 2:

COMMUNICATION



Oxidation of 5-methylcytosine: Tet family proteins can oxidize 5-methylcytosine (mC) to 5-hydroxymethylcytosine (hmC). The sequence specificity of mC oxidation by Tet1 protein was examined using randomized DNA pool and high-throughput sequencing.

*Seiichiro Kizaki, Anandhakumar Chandranan, Hiroshi Sugiyama**

Page No. – Page No.

Identification of sequence specificity of 5-methylcytosine oxidation by Tet1 protein with high-throughput sequencing

Supplementary Information

Identification of sequence specificity of methylcytosine oxidation by Tet1 protein with massively parallel sequencing

Seiichiro Kizaki,[a] Anandhakumar Chandran[a] and Hiroshi Sugiyama[a,b]

Oligomer sequences used in Bind-n-Seq method (referenced in experimental section in main text)

Bind-n-Seq 92 mer DNA:

5'-
CCATCTCATCCCTGCGTGTCTCCGACTCAGBBBBBBBBBBNNNNmCNNNNNNATCACCGACTGCCCATAGAGAGGAAAGCGGAGGCGTAGTGG-3'

• Ion A1 adapter - CCATCTCATCCCTGCGTGTCTCCGACTCAG

• Ion P1 adapter – ATCACCGACTGCCCATAGAGAGGAAAGCGGAGGCGTAGTGG

• Ion semiconductor supportive barcode region - BBBBBBBBBB , 10-letter barcodes used to separate sequencing reads for various samples as per Ion torrent sequencing technologies.

• All the barcoded Tet oxidative substrate encompassing 92 mer ssDNA library was synthesised using machine mixing followed by standard desalting purification by Sigma Aldrich.

PE primer:

5'-CCA CTA CGC CTC CGC TTT CCT CTC TA-3'

• Primer extension reaction was performed using the PE primer.

• Synthesized by Sigma Aldrich, purification by standard desalting.

Table S1 Relative enrichment of 9-base pair "k"mers. (only top 10 sequences are shown). Randomized DNA were incubated with 6.63 μM TET protein. C in center is mC in original DNA sequence and hmC in the enriched sequence.

Rank	Sequence	Enrichment
1	TCTA**C**GCTG	47
2	TCTA**C**GCTT	43
3	GATC**C**TACT	42
4	TCCT**C**CCTT	41
5	TCCG**C**TTGC	39
6	TCCG**C**TTCC	39
7	TCTG**C**CCTT	38
8	TCCG**C**TGCT	37
9	TTAC**C**AGCT	36
10	TCAT**C**GCTT	35

Table S2 Relative enrichment of 9-base pair "k"mers. (only top 10 sequences are shown). Randomized DNA were incubated with 0.663 μM TET protein. C in center is mC in original DNA sequence and hmC in the enriched sequence.

Rank	Sequence	Enrichment
1	ACTC**C**GTGC	32.596
2	TCGA**C**CTTC	32.596
3	GAGC**C**GTGT	30.622
4	GTGA**C**TATC	28.647
5	GTGA**C**GGAT	28.647
6	TATC**C**TGAC	28.647
7	TGCT**C**CTGC	28.647
8	AATG**C**ATTC	27.659
9	TGCT**C**TCGA	27.659
10	GATT**C**GATG	27.659

Fig. S1 Pulldown of hmC-containing 6-mer DNA using magnetic beads coated with anti-hmC antibody.

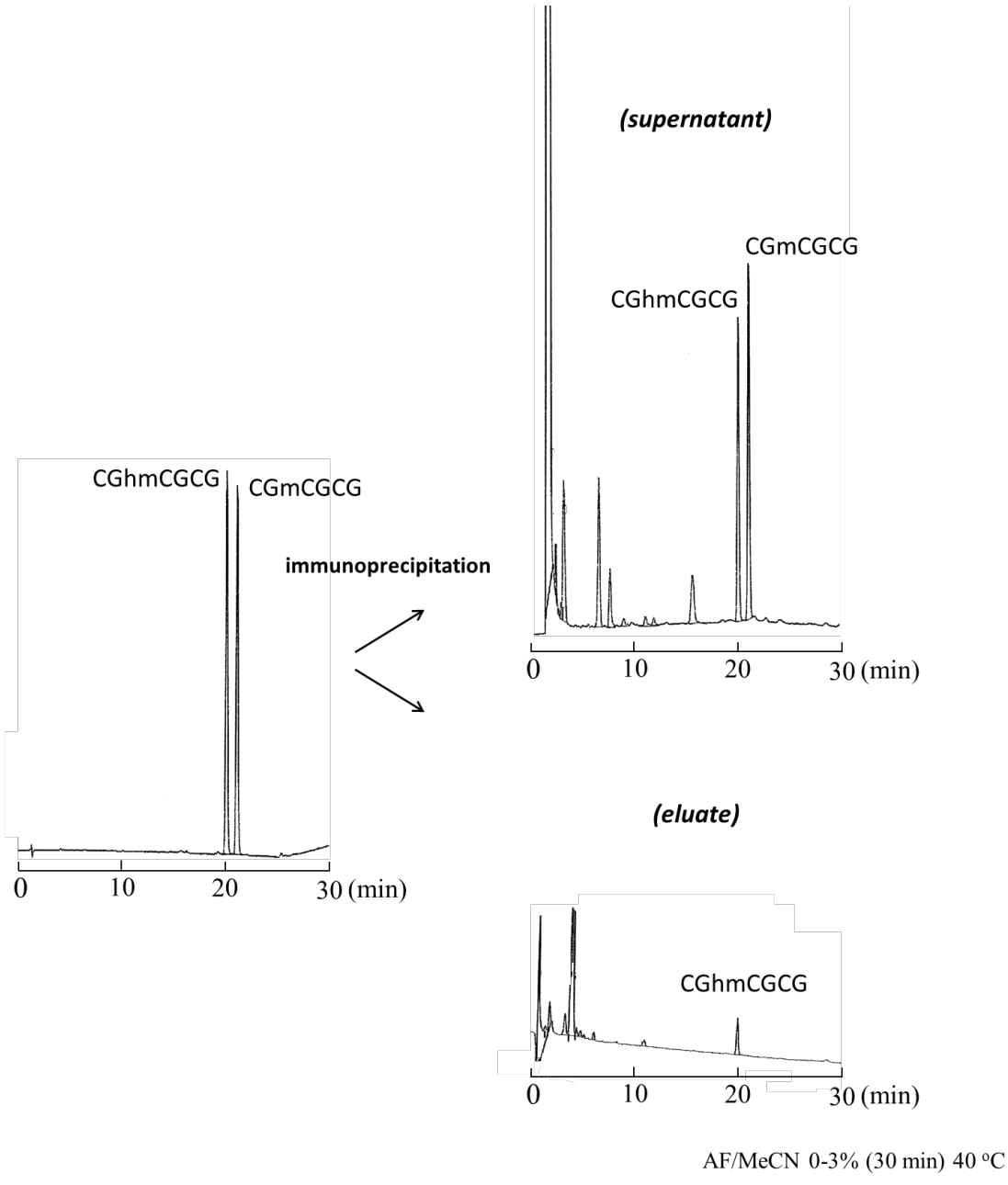


Fig. S2 Intermediate motifs identified using Bind-n-Seq through next generation sequencing. C in center is mC in original DNA sequence and hmC in the enriched sequence.

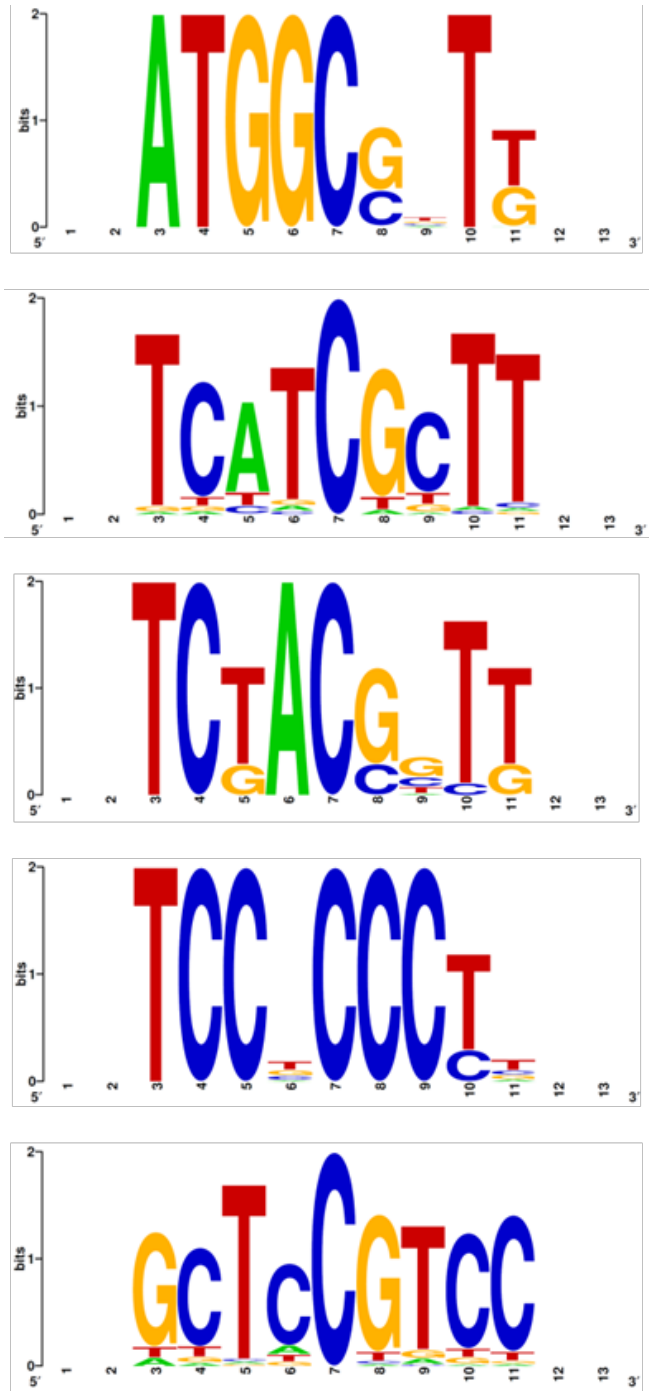
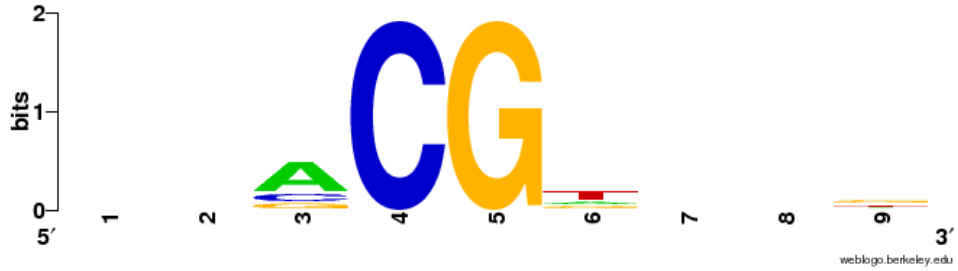


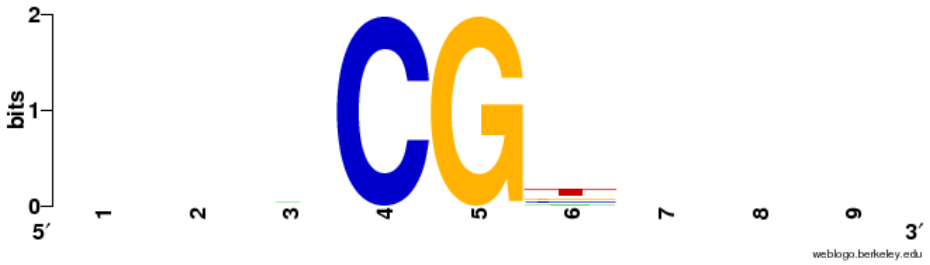
Fig. S3 Motifs derived from the hmC specific SMART sequencing of mESC genome. Sequence data were obtained from reference 1.

hmC identified in (+) strand



hmC identified in (-) strand

hemi-hmC identified in (+) strand



Hemi-hmC identified in (-) strand

References

1. C.-X. Song, T. a Clark, X.-Y. Lu, A. Kislyuk, Q. Dai, S. W. Turner, C. He, and J. Korlach, *Nat. Methods*, 2012, **9**, 75–7.