npg

LETTER TO THE EDITOR

Base-resolution maps of 5-formylcytosine and 5-carboxylcytosine reveal genome-wide DNA demethylation dynamics

Dear Editor,

The TET family of dioxygenases can oxidize 5-methylcytosine (5mC) to 5-hydroxymethylcytosine (5hmC), 5-formylcytosine (5fC), and 5-carboxylcytosine (5caC) in mammalian genomic DNA via a stepwise manner [1-5]. 5fC and 5caC are selectively recognized and excised by mammalian thymine DNA glycosylase (TDG), and restored to normal cytosine through base excision repair [3, 6-9]. Once converted to 5fC and 5caC, the modified cytosine base is presumably committed to demethylation through the TDG-dependent pathway or other potential mechanisms. Thus 5fC and 5caC specifically mark active demethylation in the mammalian genome.

To better understand the Tet/Tdg-mediated 5mC oxidation and demethylation, we and others have developed profiling methods for 5fC and/or 5caC, in mouse ESCs [10-12]. These studies revealed the preferential occurrence of 5fC and 5caC at low-methylated regions, active enhancers, and pluripotency TF-binding sites. However, whether and how 5fC and 5caC possess unique features associated with active demethylation, active enhancers and functional genomic elements are still unclear. Genome-wide, single-base resolution maps of 5fC and 5caC are required to reveal their roles in genome-wide DNA demethylation dynamics as well as their distinct properties.

We have demonstrated that chemical modification-assisted bisulfite sequencing (CAB-seq) can detect the base-resolution information of 5fC and 5caC [10, 13]. However, due to their low abundance, a direct, whole-genome bisulfite sequencing is impractical. We present here a pre-enrichment-based bisulfite sequencing strategy, or DNA immunoprecipitation-coupled CAB-seq (DIP-CAB-seq; Figure 1A and Supplementary information, Figure S1A), to generate the genome-wide, single-base resolution maps for 5fC and 5caC [10, 13, 14].

We first confirmed that this approach can effectively amplify and detect 5fC and 5caC signals on model DNA at single-base resolution (Supplementary information, Figure S1B-S1E). Sanger sequencing showed a high protection rate of CAB-seq from deamination (Supplementary information, Figure S1D); the overall signal is increased from 5% to 50% on the model DNA (Supplementary information, Figure S1E). The protection rate of 5fC by $EtONH_2$ is similar to that by the $NaBH_4$-mediated reduction reported by us [10] and in redBS [14] (Supplementary information, Figure S1F-S1H).

Using the DIP-CAB-seq approach, we generated both profiling and single-base resolution maps of 5fC and 5caC in $Tdg^{fl/fl}$ and $Tdg^{-/-}$ mouse ESCs with more than 65% mapping efficiency (Supplementary information, Table S1). The profiling analysis indicated that both 5fC and 5caC accumulate at enhancer regions in $Tdg^{-/-}$ mouse ESCs as well as major satellite repeats in both $Tdg^{fl/fl}$ and $Tdg^{-/-}$ mouse ESCs, which is consistent with previous findings [10, 11] (Supplementary information, Figure S1I-S1J). A substantial increase of 5caC signals but not 5fC at promoter regions with Tdg knockout was also observed (Figure 1B), indicating distinct features of 5fC and 5caC in the mammalian epigenome.

We next analyzed 5fC- and 5caC-enriched regions using the base-resolution data (Supplementary information, Figure S1K). We determined that 5fC and 5caC signals are significantly increased in $Tdg^{-/-}$ mouse ESCs compared to wild-type mouse ESCs (Supplementary information, Figure S1L). We observed 7 213 5fC sites and 4 806 5caC sites ($P < 0.05$), with only 6.53% ($n = 314$) overlapping with each other (Figure 1C). Consistent with the profiling results, a major portion of 5fC and 5caC sites are located in intragenic regions, especially in coding exons and introns (Figure 1D). A fraction of 5fC and 5caC sites reside in 5′ UTR, 3′ UTR, promoters and TTS regions. Notably, 5fC occurs more frequently in exons, promoter regions, and less frequently in intron regions than 5caC (Figure 1D). These results reveal that 5fC and 5caC may mark or represent distinct 5mC and 5hmC oxidation sites, which may have functional implications.
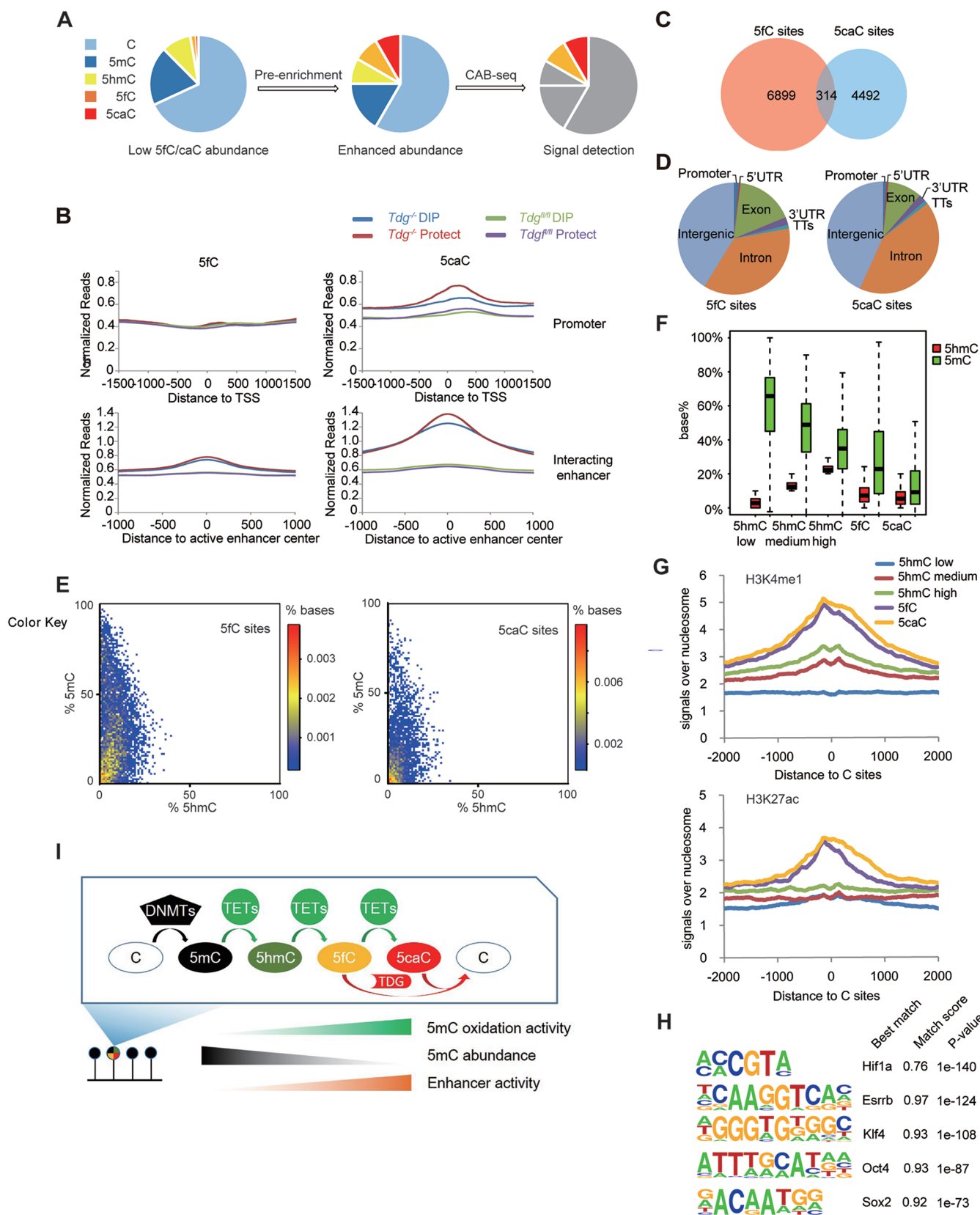
Figure 1 (A) DNA immunoprecipitation-coupled chemical modification-assisted bisulfite sequencing of 5fC and 5caC in genomic DNA. The 5fC and 5caC signals are amplified and detected by CAB-seq following pre-enrichment. **(B-D)** Generation, annotation and comparison of genome-wide profiling and base-resolution maps of 5fC and 5caC. The distribution of 5fC and 5caC signals in *Tdg^fl/fl* and *Tdg^-/-* mouse ES cells at the promoter and interacting enhancer regions, respectively **(B)**. Venn diagram presentation of the number of 5fC and 5caC sites, and their overlap **(C)**. Pie chart showing the overall distribution of 5fC and 5caC sites in genomic elements **(D)**. **(E-I)** The 5mC oxidation activity correlates with the extent of DNA hypomethylation and enhancer activity. Heatmaps of 5hmC and 5mC percentages at 5fC and 5caC sites. The occupancies of 5hmC and 5mC were estimated based on TAB-seq and traditional bisulfite sequencing data **(E)**. The levels of 5hmC and 5mC in several groups of modified cytosines enriched with 5hmC, 5fC and 5caC. 5hmC sites were further divided into low (0% - 10%), medium (10% - 20%), high (> 20%) subgroups (5 000 sites were randomly selected for each subgroup) **(F)**. The distribution of H3K4me1 and H3K27ac signals at each modified cytosine groups. ChIP-seq signals were divided by nucleosome-seq signals **(G)**. *De novo* motif analysis by HOMER [15] at ±100 bp region around 5caC sites **(H)**. Schematic diagram illustrating the dynamic equilibrium between DNMT-based cytosine methylation and Tet/Tdg-mediated 5mC oxidation and active demethylation. A gradient of 5mC oxidation activity to 5hmC, 5fC and 5caC, correlates with decreased DNA methylation level and increased enhancer activity at regulatory elements **(I)**.

By combining the single-base resolution 5fC and 5caC maps with 5mC and 5hmC maps, we plotted the percentages of 5mC and 5hmC at each 5fC and 5caC sites, respectively (Figure 1E). The abundance of 5mC notably decreases at 5caC sites compared to 5fC sites, associating the Tet-mediated 5mC oxidation with hypomethylated regions. We then sectionalized the Tet oxidation level by dividing the 5hmC sites into three sets with low (0%-10%), medium (10%-20%) and high (> 20%) 5hmC percentage (Figure 1F). The median 5mC abundance gradually decreases from 65.75% to 34.92%, accompanied by the increase of 5hmC abundance and the appearance of 5fC and 5caC sites, suggesting that lower methylated regions correlate with higher demethylation activities. Although the decrease of the 5mC and 5hmC abundance in 5fC-marked regions has been noticed [10, 11], our single-base resolution analysis demonstrated the close correlation between the extent of hypomethylation and 5mC oxidation.

To further characterize 5fC and 5caC sites, we calculated the ChIP-seq signals of enhancer histone modification markers at these regions (Figure 1G). We observed higher H3K4me1 signals at 5fC sites and even higher at 5caC sites when compared to 5hmC sites. In comparison, H3K27ac, the active enhancer marker, exhibits weak signals at 5hmC sites, but much higher signals at 5fC and 5caC sites with the highest signals observed at 5caC sites. These results indicate a gradient of Tet-mediated 5mC oxidation activity at enhancer sites that is positively correlated with enhancer activity and the extent of hypomethylation. Indeed, when we evaluated the overlap between enhancers and the modification sites, we observed an increased association percentage of these sites within interacting enhancers in the order of 5caC > 5fC > 5hmC (Supplementary information, Figure S1M-S1N). Accordingly, we observed an increased Tet1 occupancy at sites in the order of 5caC > 5fC > 5hmC (Supplementary information, Figure S1O).

Taken together, our results reveal that 5hmC, 5fC and 5caC mark distinct genomic elements with 5fC and 5caC representing more active markers compared to 5hmC. A gradient of 5hmC, 5fC and 5caC exists in genomic elements, which represents reduced DNA methylation and increased enhancer activity.

We aligned 5fC and 5caC sites in CG context and examined the base compositions (Supplementary information, Figure S1P). On the strand containing 5fCG, we observed higher percentages of cytosine than the strand containing 5mCG, 5hmCG or 5caCG, while 5caC possesses a similar local sequence context to 5hmC. Within the 100 bp window, we found that thymine composition shows a significant asymmetric distribution for 5fC, indicating that 5fC and 5caC locate in different sequence context regions, which may play a functional role in the recruitment of Tet proteins.

We then performed *de novo* motif analysis with HOMER [15] at ±100 bp region around the 5fC and 5caC sites. Consistent with the previous profiling results [11], we identified the Klf4, Oct4, Hif1a, Esrrb, and Sox2 motifs around the 5caC sites in *Tdg^-/-* mouse ESCs (Figure 1H). Using our single-base resolution map, we observed that 5caC is enriched in the Esrrb-, Klf4-, Oct4-, and Sox2-binding sites, while 5fC is evenly distributed at lower levels in these regions, suggesting high Tet oxidation activity of these sites (Supplementary information, Figure S1Q). These results indicate that Tet-based 5mC oxidation activity is the highest at these regions in order to ensure activation of these regulatory elements (Figure 1H).

In summary, we apply a pre-enrichment-based strategy for single-base resolution detection of 5fC and 5caC in mouse ESCs. We reveal a genome-wide gradient of 5mC oxidation activity at regulatory elements, which positively correlates with enhancer activity and negatively correlates with 5mC abundance (Figure 1I). Both 5fC and 5caC mark more active regulatory elements than 5hmC.

The 5caC sites represent the most active enhancers among sites with Tet-mediated modifications. At base resolution, 5fC and 5caC exhibit very limited overlap, suggesting their distinct roles even though they are all derived from oxidation of 5mC and 5hmC and excised by TDG in active demethylation. Our study reveals a highly orchestrated 5mC and 5hmC oxidation and demethylation to balance genome-wide methylation/demethylation dynamics in order to accomplish epigenetic regulation at different functional elements.

Acknowledgments

Xingyu Lu[1, *], Dali Han[1, *], Boxuan Simen Zhao[1, *], Chun-Xiao Song[1], Li-Sheng Zhang[1], Louis C Doré[1], Chuan He[1]

[1]*Department of Chemistry and Institute for Biophysical Dynamics, Howard Hughes Medical Institute, The University of Chicago, 929 E 57th Street, Chicago, IL 60637, USA*
*These three authors contributed equally to this work.
Correspondence: Chuan He
E-mail: chuanhe@uchicago.edu

References

1 Kriaucionis S, Heintz N. *Science* 2009; **324**:929-930.
2 Tahiliani M, Koh KP, Shen Y, *et al. Science* 2009; **324**:930-935.
3 He YF, Li BZ, Li Z, *et al. Science* 2011; **333**:1303-1307.
4 Ito S, Shen L, Dai Q, *et al. Science* 2011; **333**:1300-1303.
5 Pfaffeneder T, Hackner B, Truss M, *et al. Angew Chem Int Ed* 2011; **50**:7008-7012.
6 Cortázar D, Kunz C, Selfridge J, *et al. Nature* 2011; **470**:419-423.
7 Cortellino S, Xu J, Sannai M, *et al. Cell* 2011; **146**:67-79.
8 Zhang L, Lu X, Lu J, *et al. Nat Chem Biol* 2012; **8**:328-330.
9 Maiti A, Drohat AC. *J Biol Chem* 2011; **286**:35334-35338.
10 Song CX, Szulwach KE, Dai Q, *et al. Cell* 2013; **153**:678-691.
11 Shen L, Wu H, Diep D, *et al. Cell* 2013; **153**:692-706.
12 Raiber EA, Beraldi D, Ficz G, *et al. Genome Biol* 2012; **13**:R69.
13 Lu X, Song CX, Szulwach K, *et al. J Am Chem Soc* 2013; **135**:9315-9317.
14 Booth MJ, Marsico G, Bachman M, *et al. Nat Chem* 2014; **6**:435-440.
15 Heinz S, Benner C, Spann N, *et al. Mol Cell* 2010; **38**:576-589.

(**Supplementary information** is linked to the online version of the paper on the *Cell Research* website.)