Published in final edited form as:
Nat Struct Mol Biol. 2012 November ; 19(11): . doi:10.1038/nsmb.2437.

New functions for DNA modifications by TET-JBP

Yun Huang and
La Jolla Institute for Allergy and Immunology, La Jolla, California

Anjana Rao
La Jolla Institute for Allergy and Immunology, La Jolla, California

Sanford Consortium for Regenerative Medicine, La Jolla, California, and the Department of Pharmacology, University of California, San Diego, San Diego, California

Anjana Rao: arao@liai.org

Abstract

TET and JBP proteins catalyze the oxidation of methylated C bases in the mammalian genome and of the methyl group of T bases in kinetoplastid genomes, respectively. A recent study in *Nature Structural & Molecular Biology* suggests a new function of 5-methylcytosine oxidation in regulating RNA polymerase II elongation rate that is reminiscent of that of base J in transcription termination in *Leishmania*.

Methylation of cytosine at position 5 is a well-known epigenetic mark on DNA. Cytosine methylation occurs predominantly at CG sequences and is thought to be pivotal in many biological processes, including zygotic differentiation, germ cell development, X inactivation, imprinting and the silencing of parasitic DNA elements in the genome[1]. DNA methylation is dynamically altered by proteins of the TET family, 2-oxoglutarate and Fe^{2+}-dependent dioxygenases that successively oxidize 5-methylcytosine (5mC) to 5-hydroxymethylcytosine (5hmC), 5-formylcytosine (5fC) and 5-carboxylcytosine (5caC)[2–4] (Fig. 1). Methylcytosine oxidation has been linked to both passive (replication-dependent) and active (replication-independent) demethylation of DNA[2,4–7], but the role of DNA methylation and 5mC-oxidation products in gene regulation has remained unclear. In a recent issue of *Nature Structural & Molecular Biology*, Kellinger *et al.*[8] provide intriguing data that suggest a functional interplay between 5mC oxidation and the rate of transcription by RNA polymerase II (Pol II).

Through *in vitro* assays with purified yeast or mammalian Pol II, Kellinger *et al.*[8] found that the presence of 5fC and 5caC on a template DNA strand caused a substantial reduction in the rate of G incorporation from GTP at the complementary position of RNA. They assembled RNA:DNA scaffolds that contained a template DNA oligonucleotide bearing C, 5mC, 5hmC, 5fC or 5caC in a CG context at a specific site, a shorter nontranscribed strand and a complementary RNA strand that terminated just before the C or modified C (Fig. 2, top). These scaffolds were incubated with mammalian or yeast Pol II and GTP, with or without additional NTPs, and the rate of G incorporation across from the C or modified C was measured.

COMPETING FINANCIAL INTERESTS
The authors declare no competing financial interests.

In a brief incubation period of 15 s, G incorporation was considerably higher with templates containing C, 5mC or 5hmC compared to templates containing 5fC or 5caC (Fig. 2, middle and bottom). Indeed, stopped-flow methods were required to assess the kinetics of incorporation with templates containing C or 5hmC. Kinetic analysis pointed to the existence of two phases of G incorporation, occurring on time scales of seconds and minutes, respectively; the slow phase was barely observed with templates containing C, 5mC or 5hmC but was prominent for templates containing 5fC or 5caC. Although other explanations are possible, Kellinger *et al.*[8] interpret the two phases as reflecting the presence of two distinct populations of Pol II: one poised for rapid G incorporation and the second a 'paused' or back-tracked population for which the rate-limiting step for G incorporation is conversion to the poised state. When kinetic constants for the fast phase were calculated, Pol II polymerization rates for G incorporation across from 5fC and 5caC were found to be strikingly reduced, to 1–2% of those observed for C or 5hmC templates. This difference is especially notable given that modifications at the 5 position of C would not normally affect its ability to base-pair with G bases. Kellinger *et al.*[8] speculate that interaction of the formyl and carboxyl groups of 5fC and 5caC with residues on Pol II alters the position or orientation of the modified cytosine in such a way as to impair its ability to interact with incoming G.

What is the physiological relevance of these *in vitro* observations? In mouse zygotes and two- to four-cell embryos, 5fC and 5caC are present at much higher levels in the paternal compared to the maternal pronucleus, as judged by immunocytochemistry[5,6]. However, the level of transcription of endogenous genes in the paternal pronucleus is four to five times higher than that in the maternal genome, on the basis of BrUTP incorporation and immunocytochemistry[9]. Although seemingly contradictory, this finding does not necessarily run counter to the *in vitro* analyses of Kellinger *et al.*[8]; it is plausible that Pol II transcription rates in zygotes are influenced by factors other than 5fC and 5caC, for instance by chromatin modifiers recruited by TET enzymes or 5hmC.

The data of Kellinger *et al.*[8] are reminiscent of a recent study by van Luenen *et al.*[10] on the function of β-D-glucosyl-hydroxymethyluracil (base J) in *Leishmania*. Base J is found in *Leishmania*, trypanosomes and other unicellular protozoan kinetoplastid flagellates, where it constitutes a small fraction (~1% or less) of T bases in DNA[11]. Base J is produced by successive hydroxylation and glucosylation of the methyl group of T; the oxidation step is catalyzed by the J-binding proteins JBP1 and JBP2, which are members of the TET-JBP superfamily of dioxygenases[12,13]. Although there is no base J in mammalian DNA, base J, 5mC and the oxidized forms of 5mC (5hmC, 5fC and 5caC) may have related functions depending on context. Like 5mC in mammals, base J is found at telomeric repeats and other transcriptionally silent regions of the kinetoplastid genome; in many cases, its presence at sites of gene expression is associated with gene silencing[11]. Specifically, to evade the immune system of its mammalian hosts, the parasite *Trypanosoma brucei* periodically switches its surface coat, which is mainly composed of variant surface glycoproteins (VSGs)[14]. The genome of *T. brucei* contains ~20 subtelomeric copies of VSG genes, of which only one is expressed and active at any given time; notably, base J is found at the ~19 inactive VSG genes but is absent from the active gene[15]. Van Luenen *et al.*[10] now report that in *Leishmania*, the small fraction (~1%) of base J that is not in telomeric repeats is located at transcription termination sites, especially where two polycistronic transcription units, transcribed in opposite directions, use a single convergent termination site. Loss of base J results in massive read-through transcription at these sites, which suggests that base J regulates Pol II–mediated transcription by stalling Pol II or otherwise specifying transcriptional termination. In this respect, base J exhibits somewhat similar properties as 5fC and 5caC, which, rather than stalling Pol II completely, greatly decrease the rate of Pol II-mediated transcription[8].

5fC has also been reported to decrease the rate of replication of plasmid DNA. 5fC and 5caC were originally thought to be oxidative DNA-damage products of 5mC. Indeed, Kamiya *et al.*[16] reported, over a decade ago, that when DNA was aerobically treated with Fenton-type reagents, the major oxidation product of 5mC was 5fC. The same group later showed that 5fC-containing plasmids replicated less efficiently than unmodified plasmids in COS-7 cells[17]. This study parallels that of Kellinger *et al.*[8] by showing a functional change in the replication efficiency of DNA containing 5fC, even though it is now known that 5fC is a natural component of the mammalian genome.

Of the three oxidized forms of 5mC generated by TET proteins, 5hmC is the most abundant (~4×10^6–6×10^6 5hmCs per diploid genome in mouse embryonic stem (ES) cells); 5fC and 5caC are present in much lower amounts (1×10^4–6×10^4 and ~1×10^3–9×10^3 in ES cells, respectively)[2–4,8]. The low levels of these modifications are likely to reflect the fact that both bases can be excised by thymine DNA glycosylase (TDG) and replaced by cytosine through base excision repair[4,18–20], a process that would effectively reverse DNA C methylation in an active, replication-independent manner (Fig. 1). To determine how the residual (unrepaired) 5fC and 5caC are coupled to transcriptional regulation, it will be necessary to develop methods to profile the genomic distribution of modified Cs, preferably at single-base resolution.

The genomic distribution of 5hmC has been profiled in ES cells by several groups using a variety of methods[21–26]. These studies showed that 5hmC is enriched at transcription start sites and within gene bodies, especially exons, as well as at enhancers and sites of transcription factor binding. There is also a strong enrichment at transcription start sites bearing both trimethylated histone H3 Lys4 (H3K4me3) and Lys27 (H3K27me3) marks (‘bivalent promoters’). Three methods developed to profile 5hmC at single-base resolution showed that, unlike 5mC, 5hmC is asymmetrically distributed in CG dinucleotides and is enriched at CpG islands (CGIs) and nearby transcription factor–binding sites[27–29]. A method of profiling 5fC was recently reported by Raiber *et al.*[30] They reacted 5fC with aldehydereactive probe to covalently attach biotin to the functional aldehyde group of 5fC and then enriched 5fC-containing DNA fragments from mouse ES cells by using streptavidin beads. Like 5hmC, 5fC was enriched at TET1-binding sites and euchromatic regions including CGIs, exons and promoters. Enrichment of 5fC at gene promoters with CGIs correlated with higher expression of the associated gene and increased levels of the H3K4me3 ‘active’ histone mark at the gene promoters. Moreover, 5fC was significantly enriched at Pol II–bound genomic regions. Together, these data suggest a strong association of 5fC enrichment in ES cell CGI promoters with active gene transcription. Further studies will be needed to reconcile these observations with the findings of Kellinger *et al.*[8]

Several intriguing questions remain to be addressed. First, what are the mechanisms that control the levels and genomic distribution of 5fC and 5caC? Although both these modified bases can be excised by TDG[4,18], Raiber *et al.*[30] showed that 5fC distribution is only partly controlled by TDG. This is consistent with the observation that TDG-knockout ES cells—which would be expected to display a tremendous buildup of 5fC and 5caC if the cycle shown in Figure 1 applied to all cytosines in the CpG context—show only a nine-fold increase in 5caC relative to wild-type ES cells, from ~1,000 to ~9,000 5caCs[3]. This increase is minor compared with the ~30 million methylcytosine residues in the ES cell genome. Part of the discrepancy may be due to decarboxylation of 5caC: Schiesser *et al.*[31] report that ES cell lysates contain a decarboxylase activity that removes the carboxyl group of 5caC, but other mechanisms may operate as well. Second, what is the real relationship between 5fC and 5caC and transcriptional regulation? It is likely that many of the discrepancies highlighted above arise from functions that differ depending on cellular context and genomic location. An important point is that TDG—which binds tightly to 5caC[19]—may

mediate transcriptional regulation through 5caC and 5fC in a manner independent of its enzymatic activity. In addition to mediating base excision repair, TDG is known to interact with several transcription factors, including histone acetyltransferases and DNA methyltransferases[32] (Fig. 3a). Another plausible scenario is that oxidized methylcytosines, or TET proteins themselves, recruit transcription and chromatin regulators of various kinds[21,33]. Identification of TET-interacting proteins and 5hmC-, 5fC- and 5caC-binding proteins will be necessary to address these questions. Third, the ability of 5fC and 5caC to decrease the transcription elongation rate of Pol II may facilitate the interaction of Pol II with diverse transcription elongation factors, chromatin regulators, histone-modifying enzymes and factors involved in pre-mRNA splicing. Shukla *et al.*[34] showed that a DNA-binding protein, CTCF, could promote the inclusion of exons flanked by weak splice sites by reducing the rate of Pol II–mediated transcription. This effect could be inhibited by DNA methylation at CTCF-binding sites, which decreases CTCF binding. Thus, the presence of 5fC and 5caC within exons or near exon-intron boundaries could potentially alter the patterns of premRNA splicing by promoting exon inclusion or exclusion (Fig. 3b). Future experiments are needed to resolve these issues.

References

1. Ooi SK, O'Donnell AH, Bestor TH. J. Cell Sci. 2009; 122:2787–2791. [PubMed: 19657014]
2. Tahiliani M, et al. Science. 2009; 324:930–935. [PubMed: 19372391]
3. Ito S, et al. Science. 2011; 333:1300–1303. [PubMed: 21778364]
4. He YF, et al. Science. 2011; 333:1303–1307. [PubMed: 21817016]
5. Inoue A, Shen L, Dai Q, He C, Zhang Y. Cell Res. 2011; 21:1670–1676. [PubMed: 22124233]
6. Inoue A, Zhang Y. Science. 2011; 334:194. [PubMed: 21940858]
7. Nabel CS, Kohli RM. Science. 2011; 333:1229–1230. [PubMed: 21885763]
8. Kellinger MW, et al. Nat. Struct. Mol. Biol. 2012; 19:831–833. [PubMed: 22820989]
9. Aoki F, Worrad DM, Schultz RM. Dev. Biol. 1997; 181:296–307. [PubMed: 9013938]
10. van Luenen HG, et al. Cell. 2012; 150:909–921. [PubMed: 22939620]
11. Borst P, Sabatini R. Annu. Rev. Microbiol. 2008; 62:235–251. [PubMed: 18729733]
12. Iyer LM, Tahiliani M, Rao A, Aravind L. Cell Cycle. 2009; 8:1698–1710. [PubMed: 19411852]
13. Iyer LM, Abhiman S, Aravind L. Prog. Mol. Biol. Transl. Sci. 2011; 101:25–104. [PubMed: 21507349]
14. Gommers-Ampt JH, Borst P. FASEB J. 1995; 9:1034–1042. [PubMed: 7649402]
15. van Leeuwen F, et al. Genes Dev. 1997; 11:3232–3241. [PubMed: 9389654]
16. Murata-Kamiya N, et al. Nucleic Acids Res. 1999; 27:4385–4390. [PubMed: 10536146]
17. Kamiya H, et al. J. Biochem. 2002; 132:551–555. [PubMed: 12359069]
18. Maiti A, Drohat AC. J. Biol. Chem. 2011; 286:35334–35338. [PubMed: 21862836]
19. Zhang L, et al. Nat. Chem. Biol. 2012; 8:328–330. [PubMed: 22327402]
20. Nabel CS, et al. Nat. Chem. Biol. 2012; 8:751–758. [PubMed: 22772155]
21. Williams K, et al. Nature. 2011; 473:343–348. [PubMed: 21490601]
22. Wu H, et al. Genes Dev. 2011; 25:679–684. [PubMed: 21460036]
23. Huang Y, Pastor WA, Zepeda-Martinez JA, Rao A. Nat. Protoc. 2012; 7:1897–1908. [PubMed: 23018193]
24. Pastor WA, Huang Y, Henderson HR, Agarwal S, Rao A. Nat. Protoc. 2012; 7:1909–1917. [PubMed: 23018194]
25. Pastor WA, et al. Nature. 2011; 473:394–397. [PubMed: 21552279]
26. Ficz G, et al. Nature. 2011; 473:398–402. [PubMed: 21460836]
27. Song CX, et al. Nat. Methods. 2011; 9:75–77. [PubMed: 22101853]
28. Booth MJ, et al. Science. 2012; 336:934–937. [PubMed: 22539555]
29. Yu M, et al. Cell. 2012; 149:1368–1380. [PubMed: 22608086]

30. Raiber EA, et al. Genome Biol. 2012; 13:R69. [PubMed: 22902005]
31. Schiesser S, et al. Angew. Chem. Int. Edn. Engl. 2012; 51:6516–6520.
32. Cortázar D, Kunz C, Saito Y, Steinacher R, Schar P. DNA Repair (Amst.). 2007; 6:489–504. [PubMed: 17116428]
33. Kallin EM, et al. Mol. Cell. 2012 Sep 13. published online.
34. Shukla S, et al. Nature. 2011; 479:74–79. [PubMed: 21964334]

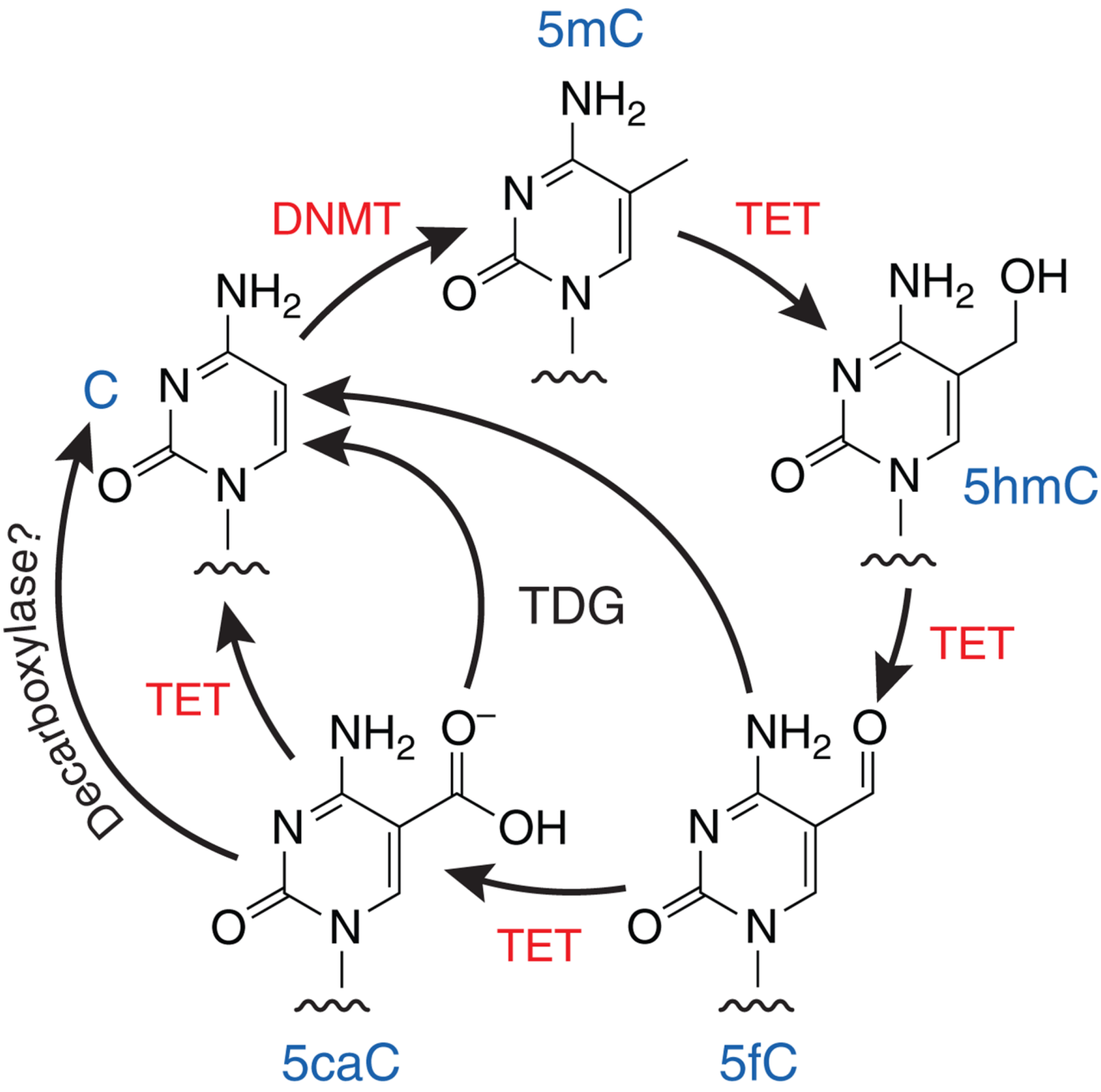

Figure 1.
The cycle of DNA methylation and demethylation by DNA methyltransferases (DNMTs) and TET proteins.

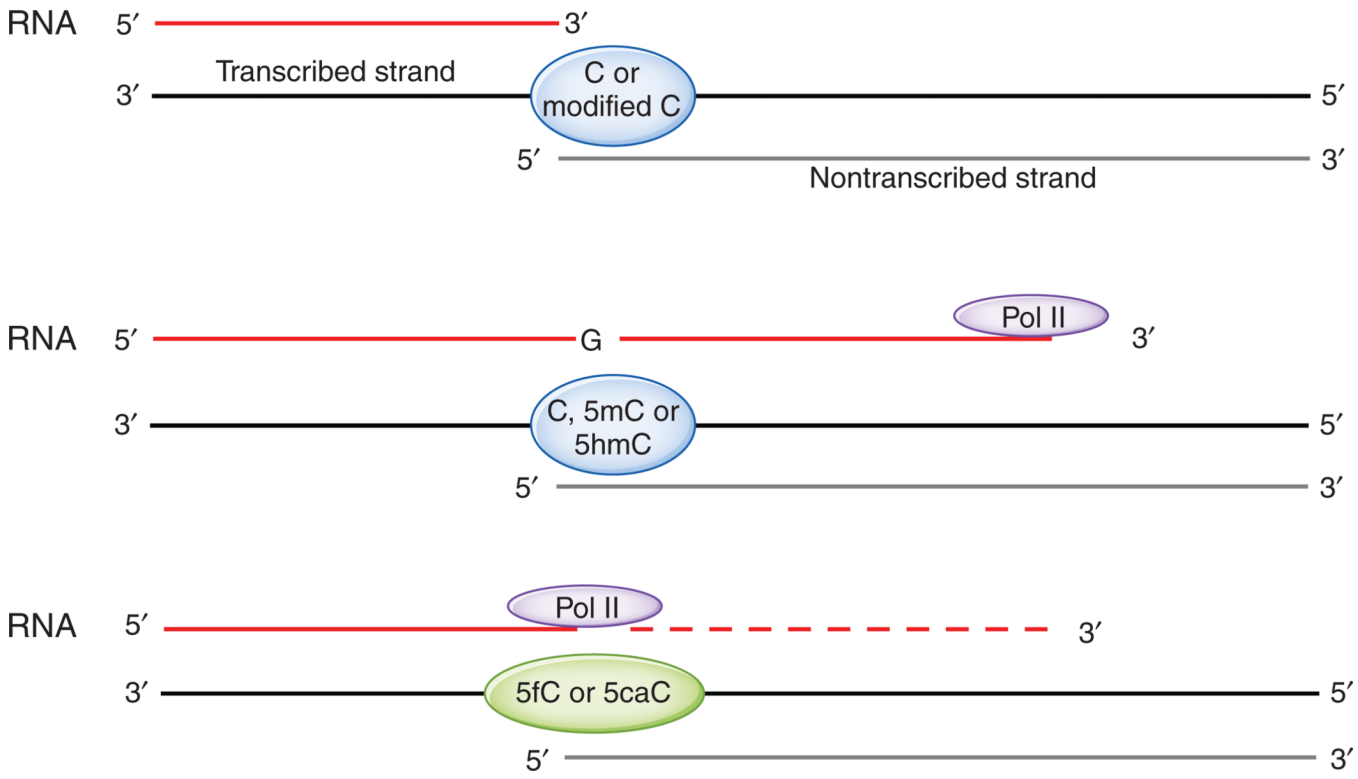

Figure 2.

The experimental setup used by Kellinger *et al*.[8] Top, one of the RNA-DNA scaffolds used in the study. Middle, Pol II can incorporate GTP into RNA across from C, 5mC and 5hmC in the transcribed template strand. Bottom, 5fC and 5caC reduce the rate of G incorporation mediated by Pol II and hence diminish Pol II processivity.

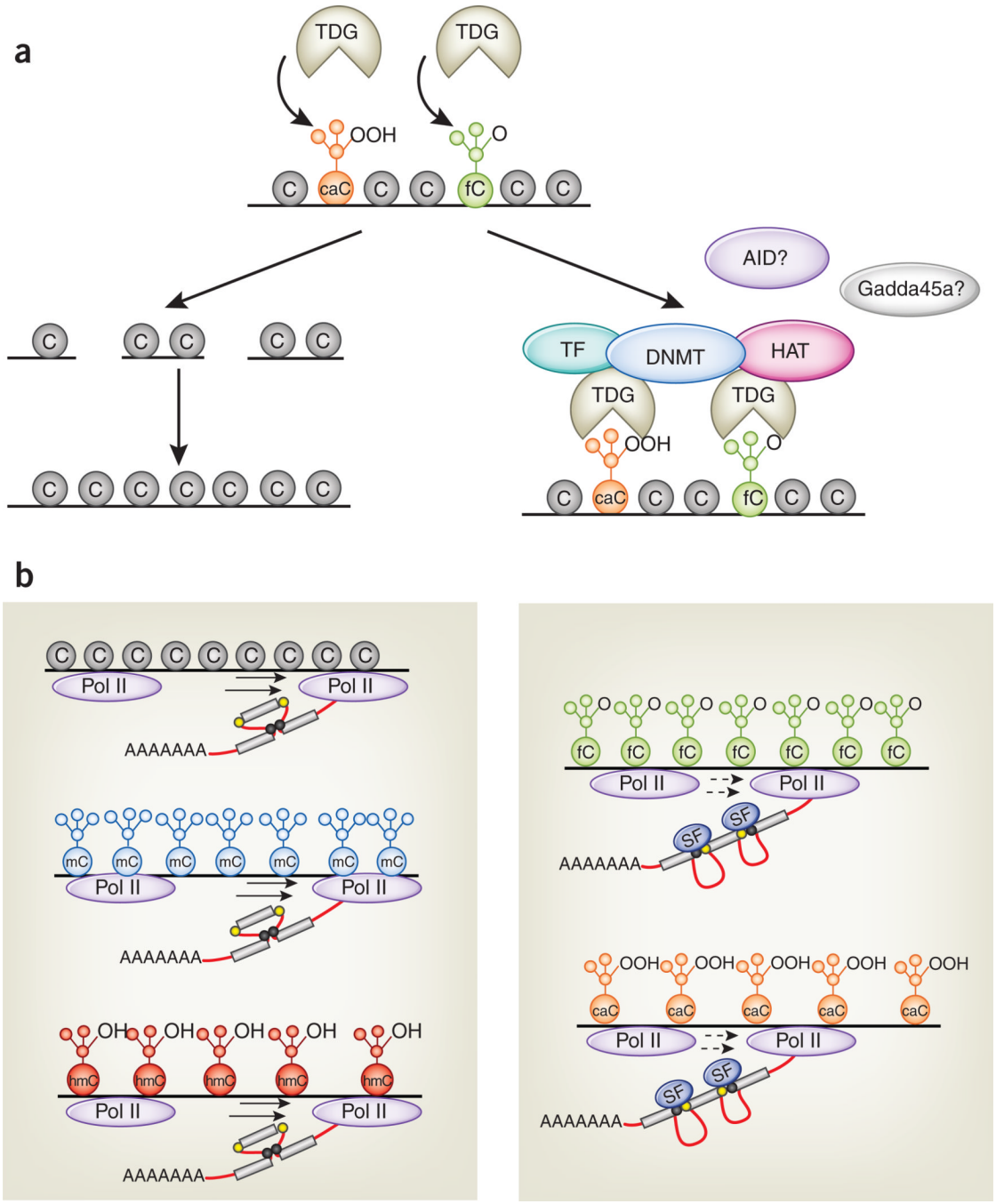

Figure 3.

Possible functions of 5fC and 5caC in transcription regulation. (**a**) 5fC (green) and 5caC (orange) are recognized by TDG and are substrates for base excision repair. TDG may also regulate transcription through recruitment of transcription factors (TF), *de novo* DNA methyltransferases (DNMT), histone acetyltransferases (HAT) and possibly the activation-induced deaminase AID and the scaffold protein Gadda45a. (**b**) 5fC and 5caC (right), but not cytosine (gray), 5mC (blue) or 5hmC (red; left), reduce the elongation rate of Pol II. One consequence of this decreased transcription rate could be to facilitate the binding of splicing factors (SF; right) to weak alternative splice sites (yellow circles), promoting the inclusion

of exons flanked by such sites. Strong splice sites not subject to this mechanism are shown as black circles in the left panel.