

This is the pre-peer reviewed version of the following article: Hardwick, J. S., Lane, A. N. and Brown, T. (2018), Epigenetic Modifications of Cytosine: Biophysical Properties, Regulation, and Function in Mammalian DNA. BioEssays, 1700199, which has been published in its final form at: www.dx.doi.org/10.1002/bies.201700199

Epigenetic modifications of cytosine: biophysical properties, regulation and function in mammalian DNA

Jack S. Hardwick ¹⁾, Andrew N. Lane ^{2)} and Tom Brown ^{1)*}*

To decode the function and molecular recognition of several recently discovered cytosine derivatives in the human genome—5-hydroxymethylcytosine, 5-formylcytosine and 5-carboxylcytosine—a detailed understanding of their effects on the structural, chemical and biophysical properties of DNA is essential. Here, we review recent literature in this area, with particular emphasis on features that have been proposed to enable the specific recognition of modified cytosine bases by DNA-binding proteins. These include electronic factors, modulation of base-pair stability, flexibility and radical changes in duplex conformation. We explore these proposals and assess whether or not they are supported by current biophysical data. This analysis is focused primarily on the properties of epigenetically modified DNA itself, which provides a basis for discussion of the mechanisms of recognition by different proteins.

¹⁾ Department of Chemistry, University of Oxford, Chemistry Research Laboratory, 12 Mansfield Road, Oxford, OX1 3TA, UK

²⁾ Department of Toxicology and Cancer Biology, University of Kentucky, 789 S. Limestone St., Lexington KY 40536, USA

***Corresponding authors:**

Tom Brown

E-mail: tom.brown@chem.ox.ac.uk

Andrew N. Lane

E-mail: andrew.lane@uky.edu

Abbreviations:

2-OG, 2-oxoglutarate; **BER**, base excision repair; **^{ca}C**, 5-carboxylcytosine; **CGI**, CpG island; **CpG**, 5'-cytosine-phosphate-guanine-3'; **DNMT**, DNA methyl transferase; **^fC**, 5-formylcytosine; **^{hm}C**, 5-hydroxymethylcytosine; **^{hm}U**, 5-hydroxymethyluracil; **IDH**, isocitrate dehydrogenase; **^mC**, 5-methylcytosine; **MBD**, methyl-binding domain; **SAM**, S-adenosylmethionine; **TDG**, thymine-DNA glycosylase; **TET**, ten-eleven translocation;

Introduction

In addition to the genetic code—based on the linear sequence of the four canonical bases of DNA: A, C, G and T—a further layer of information can be encoded into DNA via chemical modification of the bases. One such modified base, 5-methylcytosine (^mC), is found in an extraordinary variety of species, and has roles ranging from bacterial warfare to human gene regulation [1, 2]. Considering the widespread occurrence of ^mC in nature, and the base-pairing properties and three-dimensional structure of DNA (Fig. 1), it seems unlikely that methylation at the 5-position of cytosine was selected arbitrarily. Of cytosine's five potential modification sites, only C5 and C6 could be methylated without disrupting base pairing (Fig. 1a). Furthermore, in both B- and A-DNA, the methyl group of 6-methylcytosine would point towards backbone atoms, resulting in steric clashes and, consequently, deviation from normal geometry (Fig. 1c). However, in the case of C5 methylation, the methyl group extends into the major groove, avoiding such clashes. Thus, the 5-position of cytosine is uniquely positioned to tolerate methylation, and potentially other modifications, with minimal effect on DNA structure and base pairing (Figs 1a, 1c). This concept is, of course, not limited to cytosine: nature's use of 5-methyluracil—thymine—in place of uracil (U) in DNA acts as a safeguard against mutation. It helps the cell to distinguish between the correct thymine bases and uracil bases that arise spontaneously *via* deamination of C [3].

Cytosine methylation in humans

In the human genome, the most prevalent modified base is ^mC, which accounts for ~1% of all nucleobases. Cytosine methylation occurs throughout the majority of the genome and is generally associated with transcriptional repression [4], though in some cases it may actually have the opposite effect [5]. Importantly, ^mC is found primarily at CpG sites—of which 60-80% are symmetrically methylated [6]—although significant non-CpG methylation occurs in embryonic stem cells, in addition to elevated overall ^mC levels. A study comparing somatic and embryonic stem cell lines found cytosine-methylation levels of 4.25% and 5.83%, respectively, with 99.8% and 75.5% occurring at CpG sites [7]. The human genome consists of 6.47 billion base pairs in the diploid state, of which 42% are GC pairs. CpG sites would thus be expected to constitute about 4.4% of the genome. However, human-genome sequencing has revealed that CpG sites occur at only 20% of the expected frequency [8]. This underrepresentation of genomic CpG sites may be explained by the fact that ^mC has an increased propensity to convert to T *via* spontaneous deamination [9], and that most CpG sites are methylated. On the other

hand, deamination of unmethylated cytosine forms uracil (U), which is readily repaired by the cell [8]. Thus, genomic $mCpG$ sites will gradually decay to TpG unless there are specific mechanisms to counteract it.

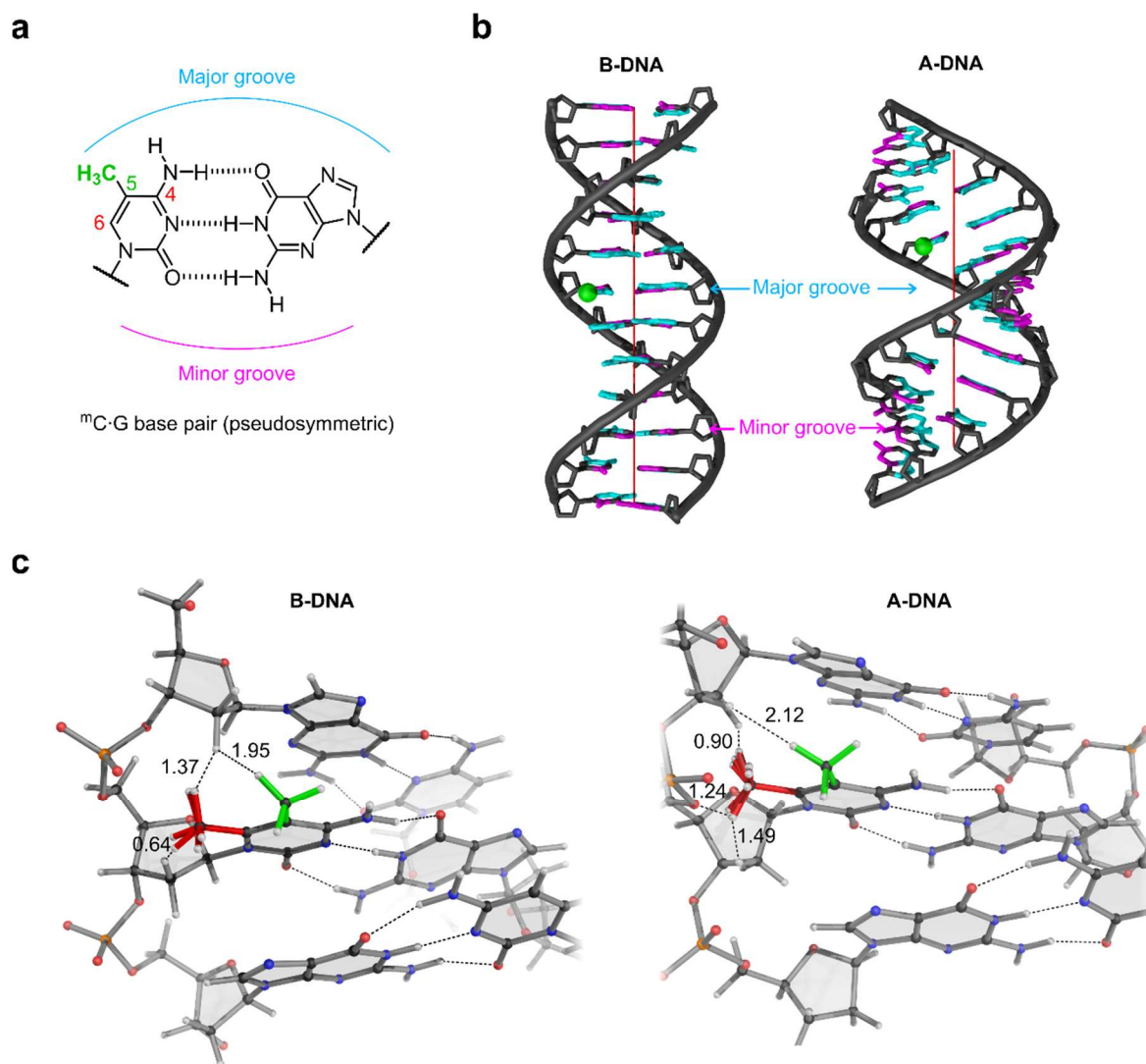


Figure 1. Cytosine methylation and DNA structure. **a)** The $mC \cdot G$ base pair. Cytosine positions 4, 5 and 6 are numbered. The 5-methyl group of mC is situated in the major groove, forming a regular pseudosymmetric base pair, and does not disrupt hydrogen bonding. Although the *anti* conformer of *N4*-methylcytosine would also not disrupt base-pairing, the *syn* conformer would. Interconversion between these states would therefore significantly lower the thermal stability of its base pair with guanine, as has been observed [10]. **b)** Models of B- and A-DNA, the two major conformational classes of DNA that are known to be biologically relevant. The major and minor groove sides are colored in cyan and magenta, respectively, the helix axes are traced in red, and the 5-position of cytosine is depicted as a green sphere. B-DNA is the predominant form in solution and in nucleosomes [11, 12], while the A-type conformation is adopted by DNA-RNA duplexes formed during transcription and, though rarely observed in solution, is frequently adopted in DNA crystal structures, due in part to the dehydrating conditions required for crystallization. **c)** B- and A-DNA models of ideal geometry showing a possible structural reason why methylation at position 5 (shown in green) of cytosine is favored over position 6 (shown in red). Distances (in Ångströms) are the minimum values obtained following rotation of the methyl groups. In both B- and A-conformations, which have different

base-stacking geometries, the presence of 6-methylcytosine would result in a number of close contacts with backbone atoms, and may therefore cause significant deviation from ideal geometry. Such deviation from the B- and A-forms could be deleterious in terms of duplex stability and during the processes of replication and transcription. Conversely, when at the 5-position, the methyl group sits comfortably in the major groove, indicating minimal impact on structure. Models were constructed using w3DNA [13] and USCF Chimera [14].

CpG islands and the role of 5-methylcytosine

Despite the underrepresentation of CpG sites in the genome, more than half of vertebrate gene promoter regions contain ~1 kb regions possessing a higher GC content (generally >50%) and increased CpG frequency [2]. These regions, known as CpG islands (CGIs), are typically hypomethylated. In fact, the very existence of CGIs is thought to be owed to their hypomethylation in the germline, whose CpG sites would otherwise have become depleted for the reasons outlined above [15]. Methylation of CGIs is associated with long-term, stable transcriptional repression and plays a fundamental role in cell differentiation [15], X-chromosome inactivation and genomic imprinting [16], though precisely how this is achieved remains unclear [2, 17]. However, the presence of ^mC can also be deleterious. Aberrant methylation is closely linked to a variety of diseases including cancer [18, 19]. Additionally, the presence of ^mC also generally increases the risk of point mutations due to its higher rate of deamination compared to C [9]. Consequently, the processes that govern the methylation and demethylation of C are vitally important, and are subjects of much ongoing research [1, 6, 20–23].

The introduction and maintenance of cytosine methylation

Cytosine methylation is catalyzed by the DNA methyl transferase (DNMT) enzymes, using SAM as the methyl donor. Three DNMT isoforms are known to regulate methylation of the human genome directly, and have different functions based on the methylation level of a given CpG site [24], each of which can contain up to two methyl marks. DNMT3A and DNMT3B favor unmethylated CpG sites, conducting *de novo* methylation during development. DNMT1, on the other hand, preferentially methylates hemi-methylated CpG sites, and is therefore important during DNA replication for copying methylation patterns to the daughter strand, referred to as maintenance methylation [25].

DNA demethylation: passive versus active

In order to remove methylation marks, the mammalian cell is faced with an apparent problem: the chemical stability of the bond between the methyl group and cytosine C5 makes its direct enzymatic cleavage highly unfavourable [26], and there is no known mammalian

enzyme that can directly remove the methylated base [27]. However, these issues can be circumvented by passive DNA demethylation, which involves the inhibition of the maintenance methylation process, leading to the steady dilution of methylation patterns upon replication [28]. Recent research has uncovered strong evidence for a replication-independent, active DNA demethylation pathway, involving sequential oxidation of ^mC to 5-hydroxymethylcytosine (^{hm}C) [29–31], 5-formylcytosine (^fC) [32, 33] and 5-carboxylcytosine (^{ca}C) [33, 34], catalyzed by the ten-eleven translocation (TET) family of enzymes. The oxidized bases ^fC and ^{ca}C can then be excised by the DNA repair protein thymine DNA glycosylase (TDG) [35, 36], and replaced by unmodified cytosine *via* the base excision repair (BER) pathway (Fig. 2d) [37]. Importantly, C, ^mC and ^{hm}C are not substrates of TDG [35, 38]; this is in contrast to their respective deamination products, U, T and ^{hm}U [39], leading to proposals of other pathways involving active deamination coupled with base excision [40–42]. Additionally, pathways have been proposed that involve direct deformylation and decarboxylation of ^fC and ^{ca}C, respectively [43, 44]. This would be an elegant solution to the problem of demethylation, avoiding the potentially damaging single- and double-stranded breaks that may form upon excision of ^fC/^{ca}C. However, the TET-TDG pathway is the most widely supported at present [45, 46].

The TET enzymes

The TET enzymes are 2-oxoglutarate (2-OG)-dependent dioxygenases that use molecular oxygen, non-haem iron, and the co-substrate 2-OG to oxidize ^mC to ^{hm}C, ^fC and ultimately ^{ca}C (Fig. 2d) with the release of succinate [47–49]. Therefore, the regulation of ^mC levels in the genome is intimately linked to one-carbon metabolism *via* the actions of DNMT and SAM, and to central metabolism *via* 2OG. It is notable that some cancer cells express a mutated form of IDH1 or IDH2 that produce 2HG that builds up to high concentrations. This compound is a potent inhibitor of the dioxygenases including TET and certain histone demethylases [50, 51].

Are the TET products more than just cytosine demethylation intermediates?

Intriguingly, several reports indicate that, in addition to their role in demethylation, oxidized derivatives of ^mC may act as distinct epigenetic signals in their own right. They are specifically recognized by proteins involved in transcriptional regulation and chromatin remodeling [52, 53], and both ^{hm}C and ^fC are reported to be genomically stable [54–56]. Evidence for readers of oxidized derivatives of ^mC has been reviewed by Song and Pfeifer [57]. Remarkably, it appears that the TET enzymes are even capable of oxidizing thymine to 5-

hydroxymethyluracil (^{hm}U) in T·A base pairs [58, 59]. This has led to the suggestion that ^{hm}U might also have an epigenetic role [58]; in mouse embryonic stem cells, it was found to be three times more abundant than ^{ca}C at peak levels. We anticipate that recent developments in mapping ^{hm}U [60], and other modified bases, which have recently been reviewed [1], will shed further light on these exciting preliminary reports.

Enzyme recognition of different cytosine states

There are numerous ^mC-binding proteins that have various roles in chromatin architecture, structure, or dynamics [61, 62]. The MeCPT2 protein contains a small ^mC-binding domain of approximately 70 residues that recognizes a single methylated CpG dinucleotide in DNA [63]. However, the protein also binds to unmethylated CpG with a differential affinity of only ~3-fold (of $\Delta G = 2.8$ kJ/mol at 310 K) [64, 65], which is consistent with the likely free energy available from direct contacts, namely van der Waals interactions, which are enthalpic and very small, and water release, which is entropic [66]. These proteins generally “recruit” additional proteins to form large complexes, suggesting the possibility of amplification of the small differences in signal by cooperative interactions. Likewise, the affinity of TET2 for DNA containing ^mC is very similar to that for unmodified C [67].

As shown in Fig. 1, the modifications at the 5-position of cytosine are accessible and have the potential to be recognized by different enzymes; for example, the proton of unmodified C is markedly different to the negatively charged carboxylate group of ^{ca}C. However, the precise recognition features and the properties of the base modifications that determine chemical reactivity toward TET and TDG are unclear. The case of TDG is particularly striking; this enzyme recognizes and removes ^fC, ^{ca}C, and T when paired with guanine in genomic DNA, but does not excise ^{hm}C, ^mC or unmodified cytosine, despite the very high prevalence of C and ^mC over ^fC and ^{ca}C. To account for this high specificity, a simple explanation is that the active site of the enzyme is sterically and electrostatically complementary to the modified base. However, analysis of various substrates of TDG's has not revealed any such interactions that could account for both substrate recognition and cytosine exclusion [39]. Therefore, several alternative mechanisms have been proposed as follows: (i) ^fC and ^{ca}C have a propensity to form asymmetric ‘wobble’ base pairs, whose distorted geometries provide a structural basis for recognition [68]; (ii) the electron-withdrawing effect of the formyl and carboxyl substituents of ^fC and ^{ca}C, respectively, weakens base pair stability relative to C, ^mC and ^{hm}C, favoring flipping of the bases into the TDG active site [69]; (iii) the electron-withdrawing

effect reduces the stability of the base-sugar bond of ^fC and ^{ca}C, leading to enhanced cleavage [35, 39]; (iv) recognition occurs via modulation of DNA structure and/or mechanical properties [70, 71]. Thus, to elucidate the mechanism(s) by which epigenetic modifications are specifically recognized by TDG and other proteins, a clear understanding of their effects on DNA structure, dynamics, base pairing and thermal stability is needed. In the following sections, we discuss recent investigations into these properties in epigenetically modified DNA, and evaluate how they may or may not facilitate the specific recognition of different cytosine states.

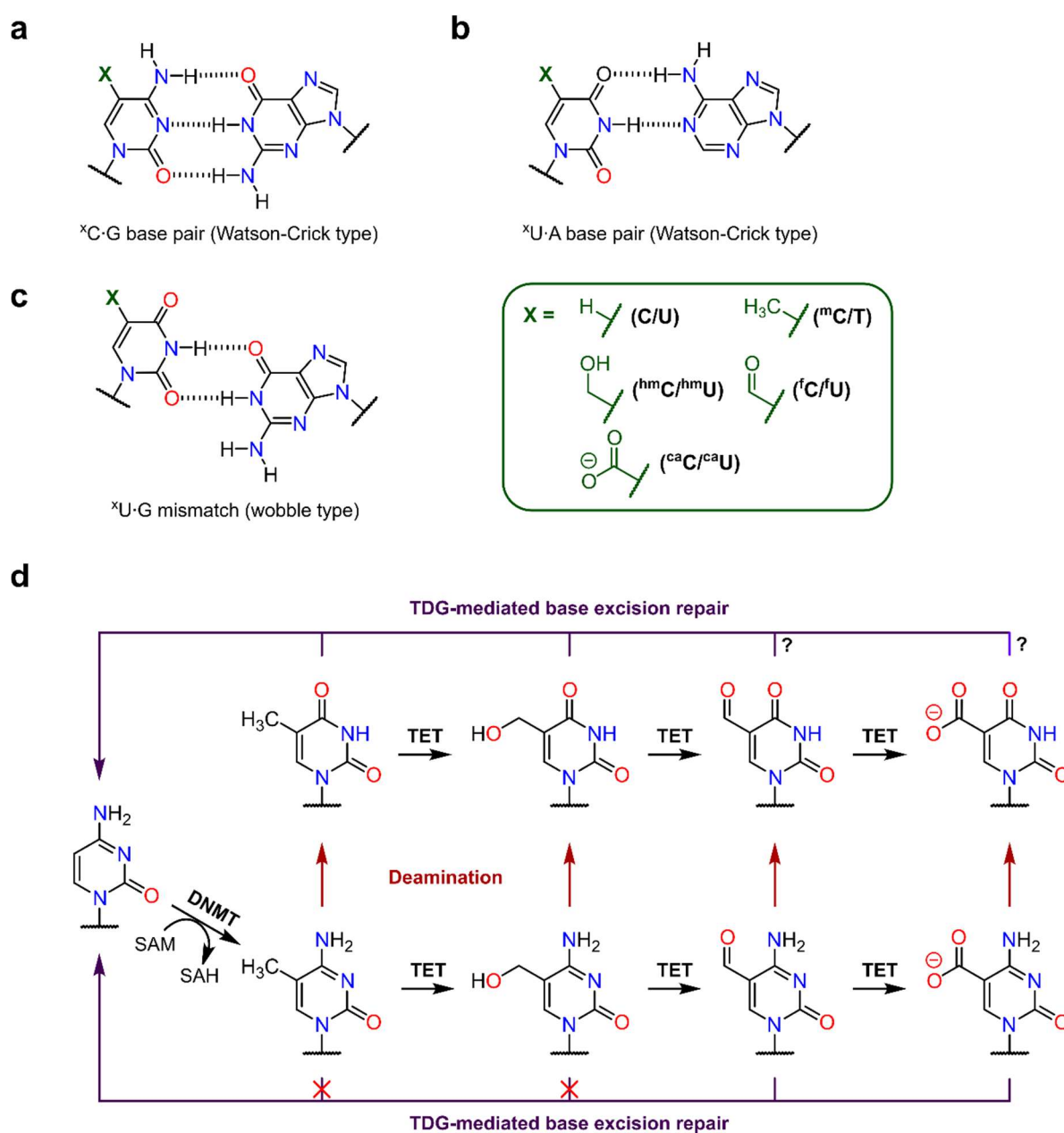


Figure 2. Cytosine modifications and their deamination products in the TET/TDG active demethylation Pathway. **a)** Watson-Crick base-pairing between cytosine derivatives and guanine. **b)** Watson-Crick base-pairing between thymine derivatives and adenine. **c)** Wobble-type mismatch formed upon cytosine deamination in C·G base pairs. **d)** Cytosine is methylated at the 5-position by the DNA methyltransferase (DNMT) enzymes, during which the cofactor S-adenosylmethionine (SAM) is converted to S-adenosylhomocysteine (SAH). ^mC can then be sequentially oxidized by the ten eleven translocation (TET) enzyme family to ^{hm}C, ^fC and finally ^{ca}C. ^fC and ^{ca}C may then be excised by thymine DNA glycosylase (TDG), which cannot remove the other cytosine derivatives. TDG can also excise the deamination products of C, ^mC and ^{hm}C (U, T and ^{hm}U, respectively). This has led to the suggestion that active DNA methylation might involve enzymatic deamination which could, in principle, obviate the need for oxidation [40–42]. However, the TET/TDG pathway is currently the most widely supported [45, 46].

Effects of cytosine modification on DNA structure and dynamics

Several crystallographic studies have been conducted in which a single addition of either ^mC, ^{hm}C, ^fC or ^{ca}C was introduced to a CpG step [72–75] in the well characterized self-complementary duplex d(CGCGAATTXGCG)₂, where X represents the modified base. All duplexes adopted the B-DNA conformation, and significant changes to global DNA conformation were not observed in any case, as shown in Fig. 3a. Differences between structures at the local level were also small, with steps containing C, ^mC, ^{hm}C, ^fC and ^{ca}C all exhibiting similar base-stacking geometries (Fig. 3b) [72–75]. Interestingly, in the structures containing ^{hm}C, the hydroxymethyl group exhibits rotational dimorphism, with two distinct conformations being refined in each crystal structure [72–74]. In every structure, the hydroxyl group was preferentially oriented in the 3' direction, within hydrogen-bonding distance of the exocyclic oxygen of G10 (Fig. 3c). Lercher *et al.* [73] noted that comparable interactions with the exocyclic heteroatoms of cytosine, thymine or adenine would not be possible, owing to geometric and distance constraints, and that therefore the presence of ^{hm}C could therefore impart a CpG-dependent effect on conformation. In the other, lower-occupancy conformation, the hydroxymethyl group appears to form a water-mediated hydrogen bond to the N7 of the 3' guanine residue. Thus, the nature of the base on the 3'-side of ^{hm}C (purine or pyrimidine) may also influence DNA conformation around ^{hm}C.

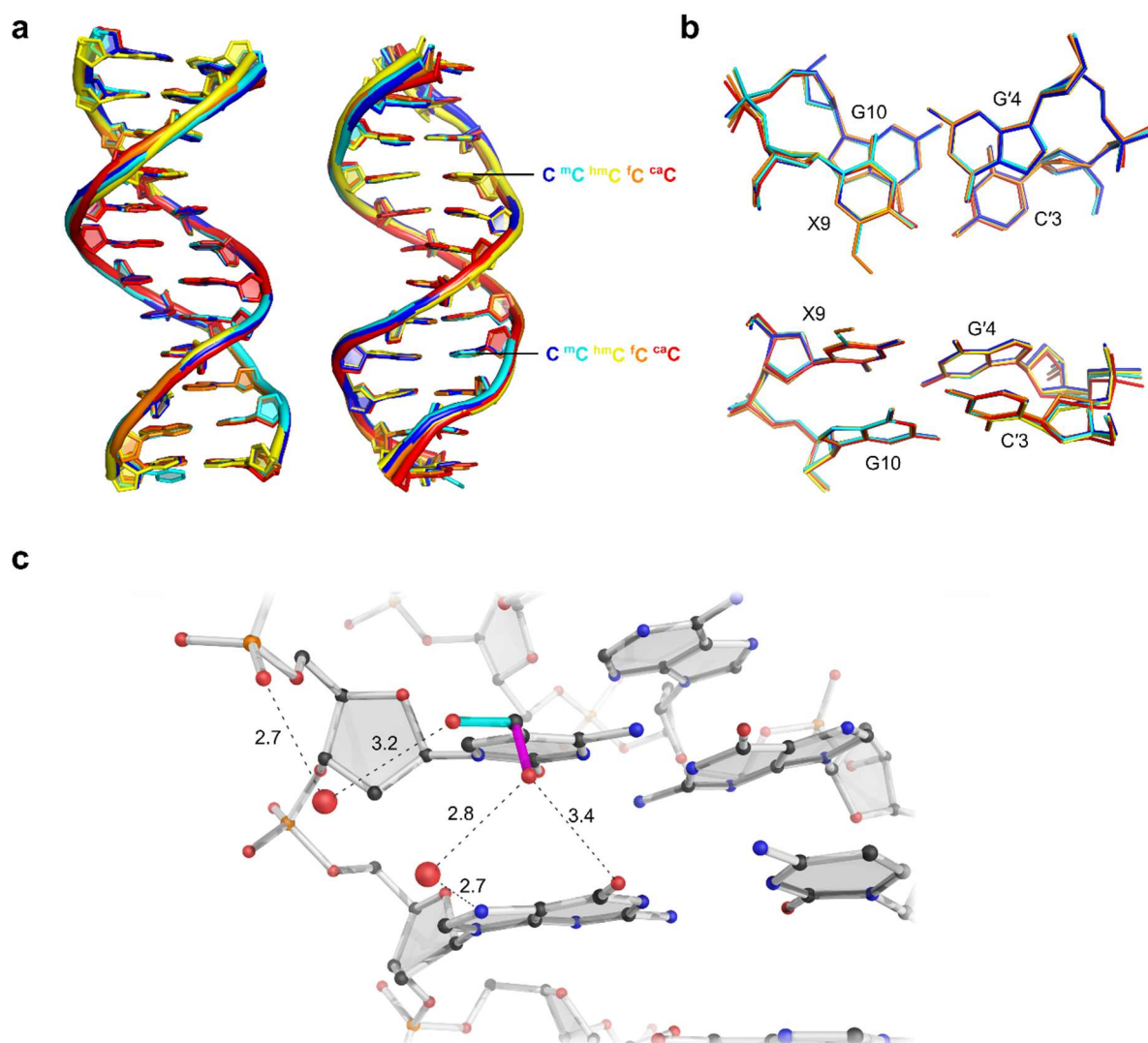


Figure 3. Comparison of crystal structures of DNA duplexes containing epigenetic cytosine modifications. **a)** Overlay of crystal structures of the duplex $d(CGCGAATT^XGCG)_2$, where X represents either C (blue, PDB 4C64), mC (cyan, PDB 4C63), ^{hm}C (yellow, PDB 4C5X), fC (orange, PDB 4QC7) or ^{ca}C (red, PDB 4PWM). **b)** Overlays showing the base-stacking between base pairs X9-G'4 and G10-C'3 of the five structures. The striking similarity in overall conformation and base-stacking geometries indicate that epigenetic cytosine modifications do not substantially perturb global or local B-DNA structure in the crystalline state. **c)** The conformational dimorphism of the hydroxyl group of ^{hm}C in the B-DNA crystal structure $d(CGCGAATT^{hm}CGCG)_2$, PDB: 4C5X. In the major conformer (magenta), the hydroxyl group can hydrogen-bond directly to the O6 of the 3' guanine in ^{hm}CpG steps, and to the N7 via a water molecule. In the minor conformer (cyan), the hydroxyl group can form a water-mediated hydrogen bond to a phosphate oxygen. The 3' guanine could thus lead to sequence-specific effects on structure and/or flexibility.

In contrast to the rotational dimorphism exhibited by the hydroxymethyl group of ^{hm}C, the formyl group of ^fC appears to be fixed in a single conformation in the plane of the cytosine ring facing towards the exocyclic N4-amine, likely forming an intramolecular hydrogen bond (Figs 4b, 5a) [48, 74–77]. NMR-based solution studies of ^fC-modified DNA duplexes [74, 76], and NMR and crystallographic studies of the free nucleoside have confirmed that this is the preferred conformation [78, 79]. Similarly, the carboxylate group of ^{ca}C also appears to form such an intramolecular hydrogen bond; crystal structures show that the carboxylate group of ^{ca}C is coplanar with the cytosine ring [74, 80, 81], and an NMR study suggests the intramolecular hydrogen bond may even be stronger than that of ^fC [74], presumably because the carboxyl group is negatively charged at physiological pH. Another potential consequence of the negative charge is that electrostatic repulsion between dicarboxylated CpG steps might lead to greater structural distortion, forcing the ^{ca}C bases further apart than the normal ~ 7.5 Å (the distance between the exocyclic carboxylate carbon atoms). However, the biological relevance of this remains to be established.

Overall, these structural analyses give no indication of significant conformational differences resulting from the presence of the epigenetic cytosine derivatives, at least for CpG steps containing a single modification. By contrast, Raiber *et al.* [70] reported that ^fC—when present in consecutive, diformylated CpG steps—substantially alters the structure of DNA, giving rise to a unique conformation which they termed F-DNA. The authors proposed that F-DNA may act as a basis for recognition of ^fC clusters by DNA binding proteins. The reportedly distinct conformation of the ^fCpG tract was attributed to unusual base-stacking geometries of ^fCpG steps, resulting from a specific hydration pattern. However, the crystal structure of the unmodified counterpart of the ^fC-duplex was not determined in the Raiber study, preventing direct structural comparison of the sequence with and without the formyl modifications to be made. Furthermore, the duplex was crystallized from a buffer containing very high salt levels, including 1.8 M lithium sulfate. Salt-dependent distortions in DNA crystal structures are well known, particularly at the duplex ends, which are more deformable [82–87]. In a comprehensive study involving X-ray crystallography, NMR, CD and UV-vis spectroscopy, we compared the structures of the ^fC-modified and unmodified duplex under similar conditions [76]. We found no evidence for F-DNA, observing minimal structural difference between the ^fC-containing core and the unmodified control both globally and locally (Figs 4a, 4b). Interestingly, the hydration pattern reportedly responsible for the F-DNA conformation was not present in the structure of the control duplex. Despite this, the duplexes are

conformationally similar in the region containing either ^fC or C (Fig. 4a), suggesting that the hydration pattern observed in the ^fC structure does not significantly influence its conformation. Furthermore, quantitative analysis of the X-ray structures, including the F-DNA structure, showed conclusively that all non-terminal residues adopt the A-DNA conformation [76]. This can be seen in Fig. 4c, which shows that the ^fC -containing region of the F-DNA structure closely resembles a standard model of A-DNA. In contrast to the crystal structures, NMR analysis showed that both the ^fC and unmodified duplexes exist as B-DNA in solution. Importantly, NMR also confirmed that conformational differences resulting from diformylated steps are small and localized. This is consistent with crystallographic studies of B-DNA containing ^fC (Figs 3a, 3b). Thus, we deduced that ^fC does not appear to change the global conformation of DNA significantly [76]. Similarly, crystallographic studies of RNA duplexes containing ^fC found that it did not cause significant global or local structural perturbation [77].

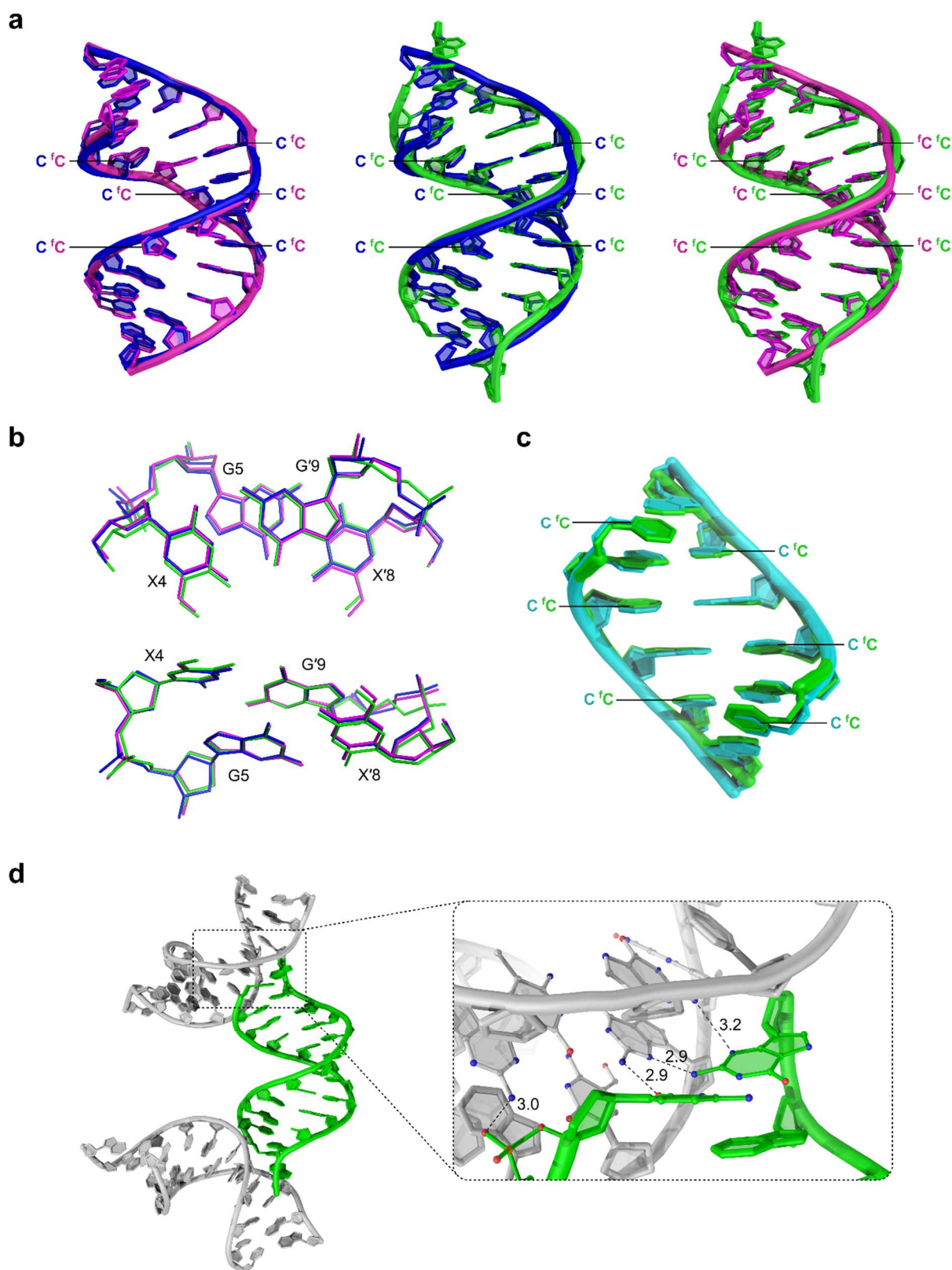


Figure 4. A crystallographic comparison of an ^{13}C -containing duplex crystallized under two different sets of conditions with its unmodified counterpart, and an A-DNA model. **a)** Overlays of crystal structures of an ^{13}C -containing duplex crystallized under two different sets of conditions and its unmodified counterpart [d(CTAXGXGXGTAG) $_2$, where X = ^{13}C or C]. The unmodified duplex (blue, PDB 5MVK) closely resembles its

⁵C-containing analogue when crystallized under moderate salt conditions (magenta, PDB 5MVU). The same ⁵C-duplex, when crystallized in a buffer containing very high salt—1.8 M lithium sulfate (green, PDB 4QKK), is also conformationally similar to the control, except at the unmodified ends, which are distorted due to crystal packing interactions [76]. **b)** Overlays showing the base-stacking between base pairs X4·G'9 and G5·X'8 of the structures featured in C, showing that the base-stacking geometry of ⁵CpG steps is not unusual. The high degree of similarity indicates that differences in hydration observed in the crystal structures are unlikely to have a significant influence on conformation. **c)** The ⁵C-containing region of the duplex, crystallized under high salt conditions, that adopts the F-DNA conformation (green), overlaid with a model of A-DNA (cyan) generated using w3DNA [13]. The high degree of overlap illustrates that the ⁵C-region of the structure adopts the A-form, which has also been confirmed quantitatively [76]. **d)** Illustration of the crystal-packing interactions of the ⁵C duplex crystallized under high-salt conditions, which are the likely cause of the structural distortion observed at the unmodified ends of the duplex. The terminal base pairs of the symmetric duplex form several close interactions with minor groove residues of neighboring duplexes in the crystal. These interactions may be stabilized by the 1.8 M lithium sulfate in the crystallization buffer, and are not seen in the crystal structure of an identical ⁵C duplex crystallized under lower salt conditions. Indeed, there are many examples in the literature of crystal-packing interactions—often salt-dependent—that lead to severe distortion of the ends of a duplex, which are more deformable than its core [82–89].

Overall, these studies suggest that any structural basis for the recognition of ⁵C is not likely to involve gross changes in DNA conformation prior to protein binding. Interestingly though, cytosine modifications have recently been reported to alter the flexibility of DNA. Ngo *et al.* found that ⁵C increases the flexibility of DNA relative to cytosine, while ⁵mC decreases it [71], suggesting that this may assist a substrate recognition mechanism for TDG through reducing the energetic penalty of bending the DNA, which is necessary to form the catalytically competent complex. However, these workers also reported increased flexibility for ⁵hmC—which is not cleaved—but did not observe increased flexibility for ⁵caC, which is cleaved. Although flexibility could be important for permitting the formation of the correct structure, it might also have a deleterious effect on the entropy of complex formation.

Effects of cytosine modification on DNA duplex stability

It has been proposed that epigenetic cytosine modifications can influence transcription [90], and other biochemical processes [69], through their modulation of base-pair stability. In general, ⁵mC appears to have a very modest stabilizing effect on C·G base pairs [72, 90–92], consistent with its minimal effect on DNA structure (see above). However, there are contradictory reports in the literature regarding the effect of the oxidized derivatives ⁵hmC, ⁵fC and ⁵caC on duplex stability, even when identical sequences were compared [69, 70]. In general, the data support the notion that hydroxylation of ⁵mCpG sites reverses the rather small stabilization afforded by ⁵mC, such that their contribution is comparable to that of unmodified CpG sites [69, 72–74, 90, 93–95]. This has led to the suggestion that ⁵hmC may, in part, reverse the transcriptionally repressive effects of cytosine methylation [90].

The effect of ^fC and ^{ca}C on duplex stability is not clear. Dai *et al.* recently reported that ^fC·G and ^{ca}C·G base pairs are significantly less stable than unmodified C·G, and proposed that this facilitates their selective excision by TDG [69]. For the sequence d(TAXGXGXGTA)₂, where X = ^fC or ^{ca}C, incorporation of ^fC caused a moderate reduction in duplex stability (3 °C – an average of 0.5 °C per ^fC residue; 10 mM phosphate buffer pH 7.5 and 0.1 M NaCl). However, ^{ca}C was found to be significantly destabilizing only in the pH range of 3-4.5 (0.1 M phosphate buffer, 0.1 M NaCl). It was argued that at low pH, protonation of the carboxylate would occur, increasing the electron-withdrawing properties of the 5-substituent and thereby weakening base pairing. At first glance, this would appear to be consistent with observations that the TDG-mediated excision of ^fC and ^{ca}C occurs by different mechanisms; the rate of ^fC excision is pH independent, while that of ^{ca}C excision increases with decreasing pH until pH 5.5, below which the enzyme denatures [96]. However, as Coey and Drohat have pointed out [45], there was no change in stability of the ^{ca}C-containing duplex at pH 5.5, despite the fact that 60% of its base pairs contained ^{ca}C [69]. Furthermore, it is not clear if the decreased stability of the ^{ca}C duplex at pH 3-5 is due to the presence of ^{ca}C, as the thermal stability of the unmodified sequence under these conditions was not reported. Given that the N3 of canonical cytosine has a pK_a of 4.5 [69], a large proportion would be expected to be protonated at pH 3, disrupting Watson-Crick base pairing and thereby strongly destabilizing the unmodified duplex. Raiber *et al.* found, using an identical sequence to Dai *et al.*, that ^fC actually increased duplex stability relative to unmodified cytosine (in phosphate-buffered saline) [70]. Others have also reported ^fC either to be weakly stabilizing [78], or to have no significant effect on duplex stability [97]. Likewise, in the case of ^{ca}C, other studies have found the modification either to be stabilizing, observing a 2-3 °C increase in melting temperature (*T*_m) per ^{ca}C residue [74, 98], or to have no significant effect [78]. Overall, the effects of epigenetic derivatives of ^mC on duplex stability remain unclear, particularly for ^fC and ^{ca}C, but generally the changes are small under near physiological conditions. They are much smaller than the destabilization caused by mismatched base pairs, and in the same range as T·A vs C·G base pairs.

It is also important to note that such studies report on global thermal stability in short oligonucleotides, where differences in duplex length, sequence context and salt concentration, nature of cations present, etc., can have a pronounced effect on melting temperature. Investigations into local stability in the context of supercoiled DNA might be more relevant, such as base pair opening [99] or bubble formation [100, 101].

Effects on base tautomerization and electronic properties

In 2001, Karino and coworkers performed single-nucleotide insertion reactions of DNA templates containing ^fC , to investigate its mutagenicity, finding that insertion of dGMP opposite to ^fC was less efficient than opposite to C, and that both dAMP and TMP were misincorporated more frequently [97]. They suggested that the intramolecular hydrogen bond between the carbonyl oxygen and the exocyclic amine may stabilize the imino tautomer, which has a different hydrogen-bonding pattern to the amino form, resembling that of thymine (Fig. 5). Thus, rather than forming the canonical pseudosymmetric base pair with guanine, the imino tautomer would favor an asymmetric, mismatch-like $^f\text{C}\cdot\text{G}$ wobble base pair, as shown in Fig. 5c. Hashimoto *et al.* have since noted that the 5-carbonyl group of ^{ca}C could also stabilize the imino form, and proposed that asymmetric, wobble-type base pairs might serve as the recognition feature for TDG substrates in general [68]. Consistent with stabilization of the imino tautomer of ^fC , Münzel *et al.* observed, using a sequencing-based mutagenicity assay, 1-2% incorporation of adenine opposite ^fC sites; however, no mutagenicity was detected for ^{ca}C [93]. Intriguingly though, $^{ca}\text{C}\cdot\text{G}$ base pairs have been reported to mimic G·T mismatches, recruiting mismatch repair proteins and promoting DNA polymerase exonuclease activity, but similar activity was not observed for ^fC [102]. Notably, biophysical studies have found no evidence for wobble base pairing of ^fC or ^{ca}C . In all duplex crystal structures, the modified bases form Watson-Crick pairs with guanine [70, 74–76], and their imino forms have not been detected by NMR [74, 76], or 2D infrared spectroscopy [69]. This is consistent with computational work by Maiti *et al.*, who calculated that the amino tautomers of ^fC and ^{ca}C are by far the most stable forms [96]. In summary, the proposal that TDG recognizes asymmetric base pairs provides a clear structural basis for TDG activity on seemingly disparate substrates T, ^fC and ^{ca}C , and also explains its rejection of C, ^mC and ^{hm}C . However, the imino tautomers of ^fC and ^{ca}C remain undetected. Nonetheless, it has been pointed out that the imino tautomer could form transiently [45], and despite being undetectable by biophysical methods, it could be biologically significant.

Bennett *et al.* have proposed another mechanism by which the electronic differences of various nucleobases may be exploited by TDG to facilitate their recognition and excision [39]. The group found a positive correlation between the leaving group ability of the nucleobase (as measured by its N1 acidity), and the rate of excision by TDG, which would simultaneously explain TDG's rejection of C, ^mC , ^{hm}C and activity towards U, T and ^fC . By contrast, the leaving group ability of ^{ca}C —which is anionic at physiological pH—is low, but its excision has been

shown to proceed via a divergent, acid-catalyzed mechanism [96]. Indeed, theoretical calculations of the three neutral forms of ^{ca}C indicate that their leaving group ability would far exceed that of the anion. This proposed mechanism has been discussed in depth in a recent review by Drohat and Coey [45]. While convincing, this mechanism does not appear to explain the fact that specific binding has been observed between a catalytically inactive TDG mutant and its known substrates [68], indicating that other factors enabling recognition must also be important.

Interestingly, we and others have observed highly unusual features in the CD spectra of various DNA duplexes containing ^fC and ^{ca}C [76]. Such dramatic changes are rarely observed in the CD spectra of DNA, particularly for relatively small nucleobase modifications. It is not surprising that such major spectral differences would be attributed to significant changes in DNA conformation, as occurred in the case of ^fC [70]. However, our NMR and UV-absorption spectroscopic studies under similar conditions to the CD studies indicate that these unusual CD features result from local electronic, rather than gross structural, differences [76]. Given the difficulty in predicting CD spectra from first principles, we are unable to interpret these effects in molecular detail. However, it is noteworthy that they occur for both cytosine derivatives that are substrates for TDG, and thus could be related to the electronic effects proposed by Bennett *et al.*, or some other special electronic features of ^fC and ^{ca}C. Theoretical studies, and further spectroscopic work, may help to explain these surprising effects.

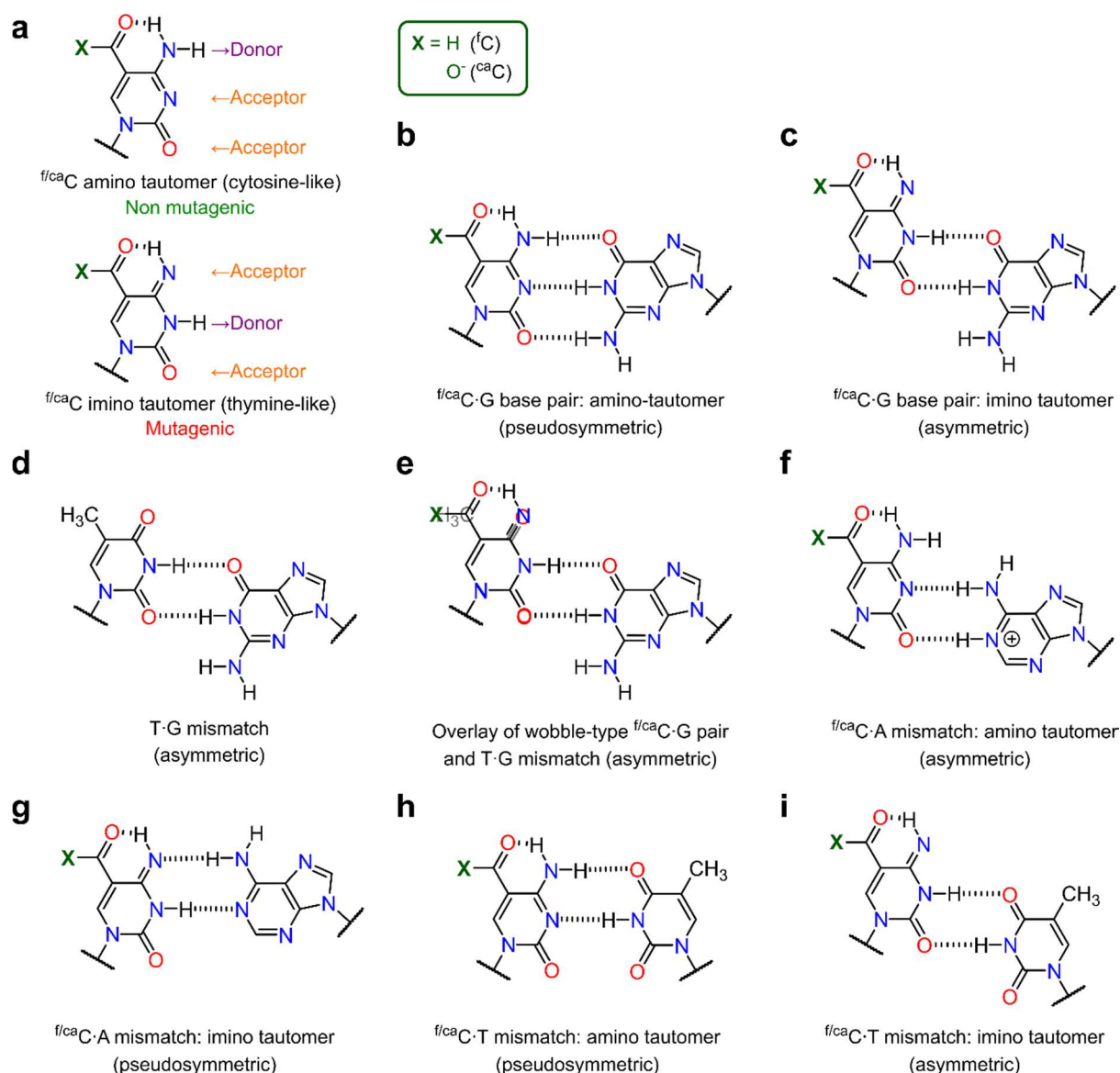


Figure 5. Skeletal structures depicting the amino (non-mutagenic) and imino (mutagenic) tautomers of $f/^{ca}\text{C}$, and their possible base pairing with G, A and T. **a)** The differing hydrogen-bonding patterns of the amino and imino tautomers of $f/^{ca}\text{C}$. The imino tautomer of $f/^{ca}\text{C}$ bears the same hydrogen-bonding pattern as thymine. **b)** The pseudosymmetric base pair formed between the amino $f/^{ca}\text{C}$ tautomer and guanine. **c)** The asymmetric ‘wobble’ base pair formed between the imino $f/^{ca}\text{C}$ tautomer and guanine. **d)** The asymmetric T·G mismatch. **e)** The T·G mismatch (grey) overlaid with the imino- $f/^{ca}\text{C}\cdot\text{G}$ base pair (black), showing their structural similarity. This asymmetric base-pairing has been proposed to act as a recognition feature for TDG, which excises ^fC , ^{ca}C and T when paired with guanine. **f)** The asymmetric amino- $f/^{ca}\text{C}\cdot\text{A}$ mismatch, which requires protonation of adenine. **g)** The imino- $f/^{ca}\text{C}\cdot\text{A}$ mismatch. Protonation of A is not required, and the base-pairing is pseudosymmetric, making its detection by repair enzymes less likely. Thus, an expected consequence of the presence of the imino tautomer of $^f\text{C}/^{ca}\text{C}$ would be mutagenesis. **h)** The amino- $f/^{ca}\text{C}\cdot\text{T}$ mismatch. Although the base pair is pseudosymmetric, it lacks shape complementarity with Watson-Crick base pairs, which would likely destabilize it relative to the imino- $f/^{ca}\text{C}\cdot\text{G}$ mismatch and make it more easily detectable. **i)** The asymmetric imino- $f/^{ca}\text{C}\cdot\text{T}$ mismatch. Overall, the imino- $f/^{ca}\text{C}\cdot\text{A}$ mismatch (g), which structurally resembles a T·A base pair, would likely be the most stable mismatch and the most likely to evade repair.

Base flipping

As discussed above, the available evidence suggests that the effects of cytosine modification on structure and thermodynamics are small, though there may be some differences in local dynamics and flexibility [72–76].

The enzymes TET and TDG must not only recognize the particular differences between unmodified and modified C, but also act on them directly, either through sequential oxidation (TET) or through base excision (TDG). Both TET and TDG locally unwind the DNA, and stabilize the state where the modified base is flipped out (Fig. 6) [48, 67, 68, 103–105].

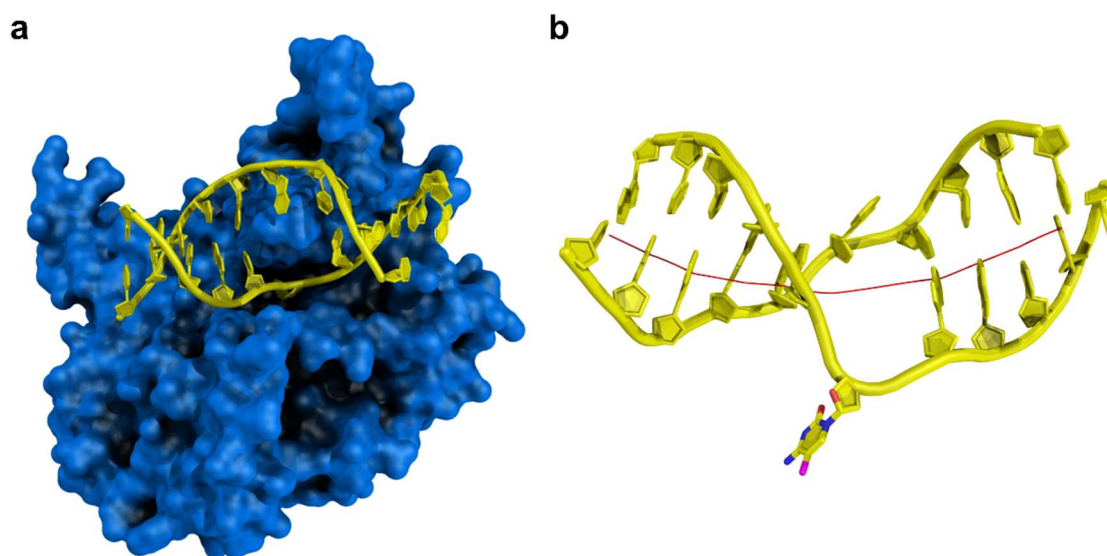


Figure 6. Crystal structure of ^mC-containing DNA in complex with TET2 (PDB: 4NM6) [67]. **a)** The entire complex. **b)** Close-up of the bound DNA. The B-DNA duplex is strongly bent and the ^mC nucleobase is flipped out into the enzyme active site, requiring the breaking of 3 hydrogen bonds and numerous stacking interactions within the double helix. This is compensated by interactions within the enzyme active site, but the methyl group itself makes no contacts. The helix axis was generated using Curves+ [106].

In the case of TET, the flipped-out C-derivative sits in the active site, and is oxidized only when there is a group available for oxidation, i.e. methyl, hydroxymethyl or formyl. Interestingly, TET is also able to oxidize T to ^{hm}U in T·A base pairs [58, 59], which implies that the enzyme does not strongly discriminate between the substituent on the 4 position of the pyrimidine ring.

Conversely, TDG will excise T in G·T mismatches, as well as ^fC and ^{ca}C in CpG context, but not C, ^mC or ^{hm}C. This selectivity is likely to be biologically important if C, ^mC and ^{hm}C are considered to be the primary epigenetic states of cytosine, as excising them is likely to be

deleterious. Thus TDG needs to discriminate C, ^mC and ^{hm}C from ^fC and ^{ca}C, as well as T·G and U·G mismatches from T·A pairs.

The flipped out base is stabilized by interactions with the enzyme at its active site. However, it is not known whether the predominant pathway is an induced fit flipping mechanism, or one in which the DNA with pre-existing flipped out pyrimidine is recognized. The spontaneous rate of flipping and the relative population of the flipped out states are likely to be small, and are certainly not normally directly observed experimentally [74, 76]. However, indirect evidence and computational studies suggest that this process does occur at slow rates that are in fact commensurate with the relevant biological timescale [107, 108]. Furthermore given the number of C residue sites in genomic DNA, even an equilibrium constant of 1×10^6 would imply more than 1000 flipped out bases at any time, and in the order of 50-100 at CpG sites. This may overestimate the numbers, as large parts of the genomic DNA are highly condensed and not accessible to these enzymes. On the other hand, the bending associated with nucleosome formation may alter the intrinsic equilibrium significantly [11, 12].

Conclusions

The last eight years have seen the discovery of three non-canonical cytosine derivatives in the human genome, and methods enabling their quantification and sequencing are now established. However, the functional roles of these modifications are still not fully understood, despite extensive efforts and significant advances. Perhaps the features recognized and acted upon by enzymes are in fact rare, transient states, captured as stable intermediates. These would be difficult to detect and characterize by methods that measure large ensemble averages—see, for example, the base-flipping of TET2 (Fig. 6). Alternatively, several features could act cooperatively to facilitate recognition. For instance, recognition and excision of substrates by TDG could exploit changes in DNA flexibility, active-site interactions and increased N1 acidity: increased DNA flexibility and active-site interactions may facilitate recognition and stabilize the reactive complex, increasing the time spent by ^fC in the flipped-out state. Once in this reactive state, electronic differences between nucleobases and specific interactions with active-site residues may stabilize the departure of the nucleobase to varying degrees, thereby governing the rate of the excision step.

Future proteomics studies may detect as-yet-unknown readers of oxidized cytosine derivatives. Such findings would not only lead to a deeper understanding of the biological

functions of these modified bases; they may also offer valuable clues into the process, or processes, of recognition itself. Interestingly, very recent work has demonstrated that reversible covalent linkages can form between the primary amines of histones and the formyl groups of ¹³C-modified DNA [109]. While the biological significance of such cross-links remains to be established, this study highlights another mechanism by which differences in chemical reactivity could be exploited to facilitate strong and specific recognition of ¹³C.

We anticipate that the coming years will provide important insights into the function and recognition of these newly-discovered cytosine derivatives.

Acknowledgments

Research of the authors is supported by an Oxford University/EPSRC Doctoral Training Partnership award (to J.S.H.), a Carmen L. Buck endowment (to A.N.L.), and a BBSRC sLoLa grant BB/J001694/2 - Extending the boundaries of nucleic acid chemistry (to T.B.).

References

1. **Raiber E-A, Hardisty R, van Delft P, Balasubramanian S.** 2017. Mapping and elucidating the function of modified bases in DNA. *Nat. Rev. Chem.* **1**: 69.
2. **Deaton A, Bird A.** 2011. CpG islands and the regulation of transcription. *Genes Dev.* **25**: 1010–22.
3. **Savva R, McAuley-Hecht K, Brown T, Pearl L.** 1995. The structural basis of specific base-excision repair by uracil–DNA glycosylase. *Nature* **373**: 487–93.
4. **Illingworth RS, Bird AP.** 2009. CpG islands - “A rough guide.” *FEBS Lett.* **583**: 1713–20.
5. **Spruijt CG, Vermeulen M.** 2014. DNA methylation: old dog, new tricks? *Nat. Struct. Mol. Biol.* **21**: 949–54.
6. **Smith ZD, Meissner A.** 2013. DNA methylation: roles in mammalian development. *Nat. Rev. Genet.* **14**: 204–20.
7. **Lister R, Pelizzola M, Dowen RH, Hawkins RD, et al.** 2009. Human DNA methylomes at base resolution show widespread epigenomic differences. *Nature* **462**: 315–22.
8. **Lander ES, Linton LM, Birren B, Nusbaum C, et al.** 2001. Initial sequencing and analysis of the human genome. *Nature* **409**: 860–921.
9. **Shen J-C, Rideout WM, Jones PA.** 1994. The rate of hydrolytic deamination of 5-methylcytosine in double-stranded DNA. *Nucleic Acids Res.* **22**: 972–6.
10. **Butkus V, Klimašauskas S, Petrauskienė L, Maneliene Z, et al.** 1987. Synthesis and physical characterization of DNA fragments containing N4-methylcytosine and 5-methylcytosine. *Nucleic Acids Res.* **15**: 8467–78.
11. **Luger K, Mäder AW, Richmond RK, Sargent DF, et al.** 1997. Crystal structure of the nucleosome core particle at 2.8 Å resolution. *Nature* **389**: 251–60.
12. **Fujii Y, Wakamori M, Umehara T, Yokoyama S.** 2016. Crystal structure of human nucleosome core particle containing enzymatically introduced CpG methylation. *FEBS Open Bio* **6**: 498–514.
13. **Zheng G, Lu X jun, Olson WK.** 2009. Web 3DNA - A web server for the analysis, reconstruction, and visualization of three-dimensional nucleic-acid structures. *Nucleic Acids Res.* **37**: 240–6.
14. **Pettersen EF, Goddard TD, Huang CC, Couch GS, et al.** 2004. UCSF Chimera: A visualization system for exploratory research and analysis. *J. Comput. Chem.* **25**: 1605–12.
15. **Jones PA.** 2012. Functions of DNA methylation: islands, start sites, gene bodies and beyond. *Nat. Rev. Genet.* **13**: 484–92.
16. **Schübeler D.** 2015. Function and information content of DNA methylation. *Nature* **517**: 321–6.
17. **Moarii M, Boeva V, Vert J-P, Reyat F.** 2015. Changes in correlation between promoter methylation and gene expression in cancer. *BMC Genomics* **16**: 873.
18. **Robertson KD.** 2005. DNA methylation and human disease. *Nat. Rev. Genet.* **6**: 597–610.
19. **Kazanets A, Shorstova T, Hilmi K, Marques M, et al.** 2016. Epigenetic silencing of tumor suppressor genes: Paradigms, puzzles, and potential. *Biochim. Biophys. Acta - Rev. Cancer* **1865**: 275–88.
20. **Carell T, Kurz MQ, Müller M, Rossa M, et al.** 2017. Non-canonical bases in the genome: The regulatory information layer in DNA. *Angew. Chemie Int. Ed.*
21. **Bergman Y, Cedar H.** 2013. DNA methylation dynamics in health and disease. *Nat Struct Mol Biol* **20**: 274–81.
22. **Lu X, Zhao BS, He C.** 2015. TET Family Proteins: Oxidation Activity, Interacting Molecules,

- and Functions in Diseases. *Chem. Rev.* **115**: 2225–39.
23. **Ooi SKT, O'Donnell AH, Bestor TH.** 2009. Mammalian cytosine methylation at a glance. *J. Cell Sci.* **122**: 2787–91.
 24. **Jones PA, Liang G.** 2009. Rethinking how DNA methylation patterns are maintained. *Nat. Rev. Genet.* **10**: 805–11.
 25. **Jin B, Li Y, Robertson KD.** 2011. DNA Methylation: Superior or Subordinate in the Epigenetic Hierarchy? *Genes Cancer* **2**: 607–17.
 26. **Kriukienė E, Liutkevičiūtė Z, Klimašauskas S.** 2012. 5-Hydroxymethylcytosine – the elusive epigenetic mark in mammalian DNA. *Chem. Soc. Rev.* **41**: 6916.
 27. **Bhutani N, Burns DMM, Blau HMM.** 2011. DNA Demethylation Dynamics. *Cell* **146**: 866–72.
 28. **Kohli RM, Zhang Y.** 2013. TET enzymes, TDG and the dynamics of DNA demethylation. *Nature* **502**: 472–9.
 29. **Kriaucionis S, Heintz N.** 2009. The Nuclear DNA Base 5-Hydroxymethylcytosine Is Present in Purkinje Neurons and the Brain. *Science (80-.)*. **324**: 929–30.
 30. **Tahiliani M, Koh KP, Shen Y, Pastor WA, et al.** 2009. Conversion of 5-Methylcytosine to 5-Hydroxymethylcytosine in Mammalian DNA by MLL Partner TET1. *Science (80-.)*. **324**: 930–5.
 31. **Ito S, D'Alessio AC, Taranova O V, Hong K, et al.** 2010. Role of Tet proteins in 5mC to 5hmC conversion, ES-cell self-renewal and inner cell mass specification. *Nature* **466**: 1129–33.
 32. **Pfaffeneder T, Hackner B, Truß M, Münzel M, et al.** 2011. The Discovery of 5-Formylcytosine in Embryonic Stem Cell DNA. *Angew. Chemie Int. Ed.* **50**: 7008–12.
 33. **Ito S, Shen L, Dai Q, Wu SC, et al.** 2011. Tet proteins can convert 5-methylcytosine to 5-formylcytosine and 5-carboxylcytosine. *Science (80-.)*. **333**: 1300–3.
 34. **He Y-F, Li B-Z, Li Z, Liu P, et al.** 2011. Tet-Mediated Formation of 5-Carboxylcytosine and Its Excision by TDG in Mammalian DNA. *Science (80-.)*. **333**: 1303–7.
 35. **Maiti A, Drohat AC.** 2011. Thymine DNA Glycosylase Can Rapidly Excise 5-Formylcytosine and 5-Carboxylcytosine: POTENTIAL IMPLICATIONS FOR ACTIVE DEMETHYLATION OF CpG SITES. *J. Biol. Chem.* **286**: 35334–8.
 36. **Raiber E-A, Beraldi D, Ficz G, Burgess HE, et al.** 2012. Genome-wide distribution of 5-formylcytosine in embryonic stem cells is associated with transcription and depends on thymine DNA glycosylase. *Genome Biol.* **13**: R69.
 37. **Weber AR, Krawczyk C, Robertson AB, Kuśnierczyk A, et al.** 2016. Biochemical reconstitution of TET1–TDG–BER-dependent active DNA demethylation reveals a highly coordinated mechanism. *Nat. Commun.* **7**: 10806.
 38. **Sibghat-Ullah, Gallinari P, Xu Y-Z, Goodman MF, et al.** 1996. Base Analog and Neighboring Base Effects on Substrate Specificity of Recombinant Human G:T Mismatch-Specific Thymine DNA–Glycosylase †. *Biochemistry* **35**: 12926–32.
 39. **Bennett MT, Rodgers MT, Hebert AS, Ruslander LE, et al.** 2006. Specificity of Human Thymine DNA Glycosylase Depends on N-Glycosidic Bond Stability. *J. Am. Chem. Soc.* **128**: 12510–9.
 40. **Morgan HD, Dean W, Coker H a., Reik W, et al.** 2004. Activation-induced Cytidine Deaminase Deaminates 5-Methylcytosine in DNA and Is Expressed in Pluripotent Tissues. *J. Biol. Chem.* **279**: 52353–60.
 41. **Popp C, Dean W, Feng S, Cokus SJ, et al.** 2010. Genome-wide erasure of DNA methylation in mouse primordial germ cells is affected by AID deficiency. *Nature* **463**: 1101–5.
 42. **Bhutani N, Brady JJ, Damian M, Sacco A, et al.** 2010. Reprogramming towards pluripotency

requires AID-dependent DNA demethylation. *Nature* **463**: 1042–7.

43. **Schiesser S, Pfaffeneder T, Sadeghian K, Hackner B**, et al. 2013. Deamination, Oxidation, and C–C Bond Cleavage Reactivity of 5-Hydroxymethylcytosine, 5-Formylcytosine, and 5-Carboxycytosine. *J. Am. Chem. Soc.* **135**: 14593–9.
44. **Globisch D, Münzel M, Müller M, Michalakakis S**, et al. 2010. Tissue Distribution of 5-Hydroxymethylcytosine and Search for Active Demethylation Intermediates. *PLoS One* **5**: e15367.
45. **Drohat AC, Coey CT**. 2016. Role of Base Excision “Repair” Enzymes in Erasing Epigenetic Marks from DNA. *Chem. Rev.* **116**: 12711–29.
46. **Wu X, Zhang Y**. 2017. TET-mediated active DNA demethylation: mechanism, function and beyond. *Nat. Rev. Genet.* **18**: 517–34.
47. **Loenarz C, Schofield CJ**. 2009. Oxygenase Catalyzed 5-Methylcytosine Hydroxylation. *Chem. Biol.* **16**: 580–3.
48. **Hu L, Lu J, Cheng J, Rao Q**, et al. 2015. Structural insight into substrate preference for TET-mediated oxidation. *Nature* **527**: 118–22.
49. **Ponnaluri VKC, Maciejewski JP, Mukherji M**. 2013. A mechanistic overview of TET-mediated 5-methylcytosine oxidation. *Biochem. Biophys. Res. Commun.* **436**: 115–20.
50. **Lu C, Ward PS, Kapoor GS, Rohle D**, et al. 2012. IDH mutation impairs histone demethylation and results in a block to cell differentiation. *Nature* **483**: 474–8.
51. **Gross S, Cairns RA, Minden MD, Driggers EM**, et al. 2010. Cancer-associated metabolite 2-hydroxyglutarate accumulates in acute myelogenous leukemia with isocitrate dehydrogenase 1 and 2 mutations. *J. Exp. Med.* **207**: 339–44.
52. **Spruijt CG, Gnerlich F, Smits AH, Pfaffeneder T**, et al. 2013. Dynamic Readers for 5-(Hydroxy)Methylcytosine and Its Oxidized Derivatives. *Cell* **152**: 1146–59.
53. **Iurlaro M, Ficiz G, Oxley D, Raiber E-A**, et al. 2013. A screen for hydroxymethylcytosine and formylcytosine binding proteins suggests functions in transcription and chromatin regulation. *Genome Biol.* **14**: R119.
54. **Su M, Kirchner A, Stazzoni S, Müller M**, et al. 2016. 5-Formylcytosine Could Be a Semipermanent Base in Specific Genome Sites. *Angew. Chemie Int. Ed.* : 1–5.
55. **Bachman M, Uribe-Lewis S, Yang X, Burgess HE**, et al. 2015. 5-Formylcytosine can be a stable DNA modification in mammals. *Nat. Chem. Biol.* **11**: 3–6.
56. **Bachman M, Uribe-Lewis S, Yang X, Williams M**, et al. 2014. 5-Hydroxymethylcytosine is a predominantly stable DNA modification. *Nat. Chem.* **6**: 1049–55.
57. **Song J, Pfeifer GP**. 2016. Are there specific readers of oxidized 5-methylcytosine bases? *BioEssays* **38**: 1038–47.
58. **Pfaffeneder T, Spada F, Wagner M, Brandmayr C**, et al. 2014. Tet oxidizes thymine to 5-hydroxymethyluracil in mouse embryonic stem cell DNA. *Nat. Chem. Biol.* **10**: 574–81.
59. **Pais JE, Dai N, Tamanaha E, Vaisvila R**, et al. 2015. Biochemical characterization of a Naegleria TET-like oxygenase and its application in single molecule sequencing of 5-methylcytosine. *Proc. Natl. Acad. Sci.* **112**: 4316–21.
60. **Kawasaki F, Beraldi D, Hardisty RE, McInroy GR**, et al. 2017. Genome-wide mapping of 5-hydroxymethyluracil in the eukaryote parasite Leishmania. *Genome Biol.* **18**: 23.
61. **Parry L, Clarke AR**. 2011. The Roles of the Methyl-CpG Binding Proteins in Cancer. *Genes Cancer* **2**: 618–30.
62. **Clouaire T, Stancheva I**. 2008. Methyl-CpG binding proteins: specialized transcriptional repressors or structural components of chromatin? *Cell. Mol. Life Sci.* **65**: 1509–22.

63. **Wakefield RI., Smith BO, Nan X, Free A,** et al. 1999. The solution structure of the domain from MeCP2 that binds to methylated DNA. *J. Mol. Biol.* **291**: 1055–65.
64. **Ishibashi T, Thambirajah AA, Ausió J.** 2008. MeCP2 preferentially binds to methylated linker DNA in the absence of the terminal tail of histone H3 and independently of histone acetylation. *FEBS Lett.* **582**: 1157–62.
65. **Hansen JC, Ghosh RP, Woodcock CL.** 2010. Binding of the Rett syndrome protein, MeCP2, to methylated and unmethylated DNA and chromatin. *IUBMB Life* **62**: 732–8.
66. **Lane AN, Jenkins TC.** 2000. Thermodynamics of nucleic acids and their interactions with ligands. *Q. Rev. Biophys.* **33**: 255–306.
67. **Hu L, Li Z, Cheng J, Rao Q,** et al. 2013. Crystal Structure of TET2-DNA Complex: Insight into TET-Mediated 5mC Oxidation. *Cell* **155**: 1545–55.
68. **Hashimoto H, Hong S, Bhagwat AS, Zhang X,** et al. 2012. Excision of 5-hydroxymethyluracil and 5-carboxylcytosine by the thymine DNA glycosylase domain: its structural basis and implications for active DNA demethylation. *Nucleic Acids Res.* **40**: 10203–14.
69. **Dai Q, Sanstead PJ, Peng CS, Han D,** et al. 2016. Weakened N3 Hydrogen Bonding by 5-Formylcytosine and 5-Carboxylcytosine Reduces Their Base-Pairing Stability. *ACS Chem. Biol.* **11**: 470–7.
70. **Raiber E-A, Murat P, Chirgadze DY, Beraldi D,** et al. 2014. 5-Formylcytosine alters the structure of the DNA double helix. *Nat. Struct. Mol. Biol.* **22**: 44–9.
71. **Ngo TTM, Yoo J, Dai Q, Zhang Q,** et al. 2016. Effects of cytosine modifications on DNA flexibility and nucleosome mechanical stability. *Nat. Commun.* **7**: 10813.
72. **Renciuk D, Blacque O, Vorlickova M, Spingler B.** 2013. Crystal structures of B-DNA dodecamer containing the epigenetic modifications 5-hydroxymethylcytosine or 5-methylcytosine. *Nucleic Acids Res.* **41**: 9891–900.
73. **Lercher L, McDonough M a, El-Sagheer AH, Thalhammer A,** et al. 2014. Structural insights into how 5-hydroxymethylation influences transcription factor binding. *Chem. Commun.* **50**: 1794–6.
74. **Szulik MW, Pallan PS, Nocek B, Voehler M,** et al. 2015. Differential Stabilities and Sequence-Dependent Base Pair Opening Dynamics of Watson–Crick Base Pairs with 5-Hydroxymethylcytosine, 5-Formylcytosine, or 5-Carboxylcytosine. *Biochemistry* **54**: 1294–305.
75. **Kimura K, Ono A, Watanabe K, Takenaka A,** X-Ray analyses of oligonucleotides containing 5-formylcytosine, suggest a structural reason for the codon-anticodon recognition of mitochondrial tRNA-Met. *To be Publ.*
76. **Hardwick JS, Ptchelkine D, El-Sagheer AH, Tear I,** et al. 2017. 5-Formylcytosine does not change the global structure of DNA. *Nat. Struct. Mol. Biol.* **24**: 544–52.
77. **Wang R, Luo Z, He K, Delaney MO,** et al. 2016. Base pairing and structural insights into the 5-formylcytosine in RNA duplex. *Nucleic Acids Res.* **44**: 4968–4977.
78. **Münzel M, Lischke U, Stathis D, Pfaffeneder T,** et al. 2011. Improved Synthesis and Mutagenicity of Oligonucleotides Containing 5-Hydroxymethylcytosine, 5-Formylcytosine and 5-Carboxylcytosine. *Chem. - A Eur. J.* **17**: 13782–8.
79. **Kawai G, Yokogawa T, Nishikawa K, Ueda T,** et al. 1994. Conformatzonal Properties of a Novel Modified Nucleoside, 5-Formylcytidine, Found at the First Position of the Anticodon of Bovine Mitochondrial tRNA Met. *Nucleosides and Nucleotides* **13**: 1189–99.
80. **Irrera S, Portalone G.** 2013. First X-ray diffraction and quantum chemical study of proton-acceptor and proton-donor forms of 5-carboxylcytosine, the last-discovered nucleobase. *J. Mol. Struct.* **1050**: 140–50.
81. **Irrera S, Ruiz-Hernandez SE, Reggente M, Passeri D,** et al. 2017. Self-assembling of calcium salt of the new DNA base 5-carboxylcytosine. *Appl. Surf. Sci.* **407**: 297–306.

82. **Liu J, Subirana JA.** 1999. Structure of d(CGCGAATTCGCG) in the presence of Ca(2+) ions. *J. Biol. Chem.* **274**: 24749–52.
83. **Liu J, Malinina L, Huynh-Dinh T, Subirana JA.** 1998. The structure of the most studied DNA fragment changes under the influence of ions: a new packing of d(CGCGAATTCGCG). *FEBS Lett.* **438**: 211–4.
84. **Valls N, Wright G, Steiner RA, Murshudov GN, et al.** 2004. DNA variability in five crystal structures of d(CGCAATTGCG). *Acta Crystallogr. Sect. D Biol. Crystallogr.* **60**: 680–5.
85. **Abrescia NG., Malinina L, Subirana J a.** 1999. Stacking interaction of guanine with netropsin in the minor groove of d(CGTATATACG)2. *J. Mol. Biol.* **294**: 657–66.
86. **Abrescia NGA, Malinina L, Fernandez LG, Huynh-Dinh T, et al.** 1999. Structure of the oligonucleotide d(CGTATATACG) as a site-specific complex with nickel ions. *Nucleic Acids Res.* **27**: 1593–9.
87. **Gao YG, Robinson H, Wang AH.** 1999. High-resolution A-DNA crystal structures of d(AGGGGCCCCCT). An A-DNA model of poly(dG) x poly(dC). *Eur. J. Biochem.* **261**: 413–20.
88. **Spink N, Nunn CM, Vojtechovsky J, Berman HM, et al.** 1995. Crystal structure of a DNA decamer showing a novel pseudo four-way helix-helix junction. *Proc. Natl. Acad. Sci.* **92**: 10767–71.
89. **Johansson E, Parkinson G, Neidle S.** 2000. A New Crystal Form for the Dodecamer C-G-C-G-A-A-T-T-C-G-C-G : Symmetry Effects on Sequence-dependent DNA Structure. : 551–61.
90. **Thalhammer A, Hansen AS, El-Sagheer AH, Brown T, et al.** 2011. Hydroxylation of methylated CpG dinucleotides reverses stabilisation of DNA duplexes by cytosine 5-methylation. *Chem. Commun.* **47**: 5325.
91. **Nardo L, Lamperti M, Salerno D, Cassina V, et al.** 2015. Effects of non-CpG site methylation on DNA thermal stability: a fluorescence study. *Nucleic Acids Res.* **43**: 10722–33.
92. **Wojdacz TK, Dobrovic A, Hansen LL.** 2008. Methylation-sensitive high-resolution melting. *Nat. Protoc.* **3**: 1903–8.
93. **Münzel M, Globisch D, Trindler C, Carell T.** 2010. Efficient Synthesis of 5-Hydroxymethylcytosine Containing DNA. *Org. Lett.* **12**: 5671–3.
94. **Carson S, Wilson J, Aksimentiev A, Weigele PR, et al.** 2016. Hydroxymethyluracil modifications enhance the flexibility and hydrophilicity of double-stranded DNA. *Nucleic Acids Res.* **44**: 2085–92.
95. **Wanunu M, Cohen-Karni D, Johnson RR, Fields L, et al.** 2011. Discrimination of Methylcytosine from Hydroxymethylcytosine in DNA Molecules. *J. Am. Chem. Soc.* **133**: 486–92.
96. **Maiti A, Michelson AZ, Armwood CJ, Lee JK, et al.** 2013. Divergent Mechanisms for Enzymatic Excision of 5-Formylcytosine and 5-Carboxylcytosine from DNA. *J. Am. Chem. Soc.* **135**: 15813–22.
97. **Karino N, Ueno Y, Matsuda A.** 2001. Synthesis and properties of oligonucleotides containing 5-formyl-2'-deoxycytidine: in vitro DNA polymerase reactions on DNA templates containing 5-formyl-2'-deoxycytidine. *Nucleic Acids Res.* **29**: 2456–63.
98. **Sumino M, Ohkubo A, Taguchi H, Seio K, et al.** 2008. Synthesis and properties of oligodeoxynucleotides containing 5-carboxy-2'-deoxycytidines. *Bioorganic Med. Chem. Lett.* **18**: 274–7.
99. **Szulik MW, Voehler M, Stone MP.** 2014. NMR Analysis of Base-Pair Opening Kinetics in DNA. In *Current Protocols in Nucleic Acid Chemistry*. John Wiley & Sons, Inc. p 7.20.1-7.20.18.
100. **Rapti Z, Smerzi A, Rasmussen KØ, Bishop AR, et al.** 2006. Lengthscales and cooperativity in DNA bubble formation. *Europhys. Lett.* **74**: 540–6.

101. **Lane AN, Chaires JB, Gray RD, Trent JO.** 2008. Stability and kinetics of G-quadruplex structures. *Nucleic Acids Res.* **36**: 5482–515.
102. **Shibutani T, Ito S, Toda M, Kanao R,** et al. 2014. Guanine- 5-carboxylcytosine base pairs mimic mismatches during DNA replication. *Sci. Rep.* **4**: 5220.
103. **Pidugu LS, Flowers JW, Coey CT, Pozharski E,** et al. 2016. Structural Basis for Excision of 5-Formylcytosine by Thymine DNA Glycosylase. *Biochemistry* **55**: 6205–8.
104. **Coey CT, Malik SS, Pidugu LS, Varney KM,** et al. 2016. Structural basis of damage recognition by thymine DNA glycosylase: Key roles for N-terminal residues. *Nucleic Acids Res.* **44**: 10248–58.
105. **Malik SS, Coey CT, Varney KM, Pozharski E,** et al. 2015. Thymine DNA glycosylase exhibits negligible affinity for nucleobases that it removes from DNA. *Nucleic Acids Res.* **43**: 9541–9552.
106. **Lavery R, Moakher M, Maddocks JH, Petkeviciute D,** et al. 2009. Conformational analysis of nucleic acids revisited: Curves+. *Nucleic Acids Res.* **37**: 5917–29.
107. **Bouvier B, Grubmüller H.** 2007. A Molecular Dynamics Study of Slow Base Flipping in DNA using Conformational Flooding. *Biophys. J.* **93**: 770–86.
108. **Ma N, van der Vaart A.** 2017. Free Energy Coupling between DNA Bending and Base Flipping. *J. Chem. Inf. Model.* **57**: 2020–6.
109. **Li F, Zhang Y, Bai J, Greenberg MM,** et al. 2017. 5-Formylcytosine Yields DNA-Protein Cross-Links in Nucleosome Core Particles. *J. Am. Chem. Soc.* **139**: 10617–20.