5-Formylcytosine alters the structure of the DNA double helix

Eun-Ang Raiber[1*], Pierre Murat[1*], Dimitri Y. Chirgadze[2], Dario Beraldi[3], Ben F. Luisi[2] & Shankar Balasubramanian[1,3,4]

[1] Department of Chemistry, University of Cambridge, Cambridge, UK.

[2] Department of Biochemistry, University of Cambridge, Cambridge, UK.

[3] Cancer Research UK Cambridge Institute, Li Ka Shing Centre, Cambridge, UK.

[4] School of Clinical Medicine, University of Cambridge, Cambridge, UK.

*These authors contributed equally to this work

Correspondence should be addressed to S.B. (sb10031@cam.ac.uk).

The modified base 5-formylcytosine (5fC) was recently identified in mammalian DNA and might be considered as the "seventh" base of the genome. This nucleotide has been implicated in active demethylation mediated by the base excision repair enzyme thymine DNA glycosylase (TDG). Genomics and proteomics studies have suggested a further role for 5fC in transcription regulation through chromatin remodeling. Herein we propose how 5fC might signal these processes through its effect on DNA conformation. Biophysical and structural analysis revealed that 5fC alters the structure of the DNA double helix leading to a conformation unique amongst known DNA structures including those comprising other cytosine modifications. The 1.4 Å resolution X-ray crystal structure of a DNA dodecamer comprising

three 5fCpG sites shown how 5fC changes the geometry of the grooves and base pairs associated with the modified base, which lead to helical under-winding.

INTRODUCTION

To date, four modified cytosines have been discovered in mammalian genomes: 5-methylcytosine (5mC), 5-hydroxymethylcytosine (5hmC), 5-formylcytosine (5fC) and 5-carboxycytosine (5caC). The discovery of these naturally occurring nucleobases has sparked the search for possible associated biological functions.[1,2] The most frequently postulated function is their role in the active DNA demethylation pathway (**Figure 1a**), a key process in re-setting epigenetic information. A vital player in this pathway is the thymine DNA glycosylase (TDG), which can excise both 5fC and 5caC, but prefers the former.[3] The mechanism, however, by which TDG recognizes the oxidized products remains unclear.[4] Recently the identification of transcriptional regulators, DNA repair factors and chromatin regulators that selectively binds to 5fC in genomic sequences, suggests that 5fC may be an epigenetic signal on its own right.[5]

As the presence of modified cytosines in mammalian genomes might have important biological consequences, we were interested in assessing the influence of modified cytosines on the thermodynamic and the structural properties of the DNA double helix. Previous reports have shown that 5mC and 5hmC does not influence either the B-DNA double helix structure or the modified base pair geometry, but increase the thermodynamic stability of the

double helix.[6,7] Due to the growing interest in 5fC function,[1,3,5,8–12] we set out to investigate the impact of 5fC on the structure of double-stranded DNA. We used a single-base resolution 5fC sequencing dataset in order to select for sequence context displaying a high level of formylation. We then performed detailed biophysical and structural analysis on the related 5fC containing DNA duplexes.

Here we show that 5fC is distinct from 5mC, 5hmC and 5caC by its pronounced impact on the structure of the DNA double helix. 5fC-containing oligonucleotides exhibited a distinct spectroscopic signature together with specific structural features found in a 1.4 Å X-ray crystal structure of a dodecamer comprising 5fC. The results presented herein provide new insights at the molecular level on how chemical modifications might impact biology.

RESULTS

Highly formylated elements are prevalent in CpG repeats

Quantitative sequencing of 5fC at single-base resolution in mouse embryonic stem cells[10] and two-cell embryos[8] reveal high level of formylated cytosine in specific genomic locations. Data extracted from 5fC sequencing of mouse two-cell embryos revealed that the highly formylated elements are found in CpG repeats ($d(CG)_n$, $n \geq 3$) (**Fig. 1b**, **Supplementary Fig. 1a -c**). We found that at such sites formylation levels of all Cs of a given CpG repeat are similar within a strand and across both strands (**Fig. 1c**, **Supplementary Fig. 1d** and **1e**), suggesting that the modifications tend to cluster. The

tendency for 5fC to occur on both strands at a modified site is consistent with recent structural and biochemical studies that show that TET enzymes preferentially oxidize 5mC in symmetric methylated CpG sites[13] and maintains symmetry of the resulting formylated CpG sites[14]. Long CpG repeats with high formylation level (up to 80%) can be observed in genes such as chromatin remodelers (e.g. *Hdac9*, *Usp22*) and transcription factors (e.g. *Maz*, *Ebf3*) (**Fig. 1d**, **Supplementary Fig. 2**). Highly formylated CpG repeats in gene bodies are preferentially found in introns (**Supplementary Fig. 3a**) and are enriched in genes associated with transcription, cell differentiation and development (**Supplementary Fig. 3b** and **3c**). Taken together, these results suggest that TET-mediated formylation of CpG repeats contributes to the regulation of gene expression and cell differentiation in mouse two-cell embryos.

Thermodynamic and spectroscopic properties of CpG repeat

In order to assess the impact of cytosine formylation within CpG repeats on the stability and structure of DNA and compare the effect of 5fC to other cytosine modifications, we prepared modified oligonucleotides whose sequences comprise CpG repeats ($d(CG)_n$, n = 3) bearing each of the known modified cytosines for biophysical analysis (ODN1-5, **Fig. 2a and b, Supplementary Table 1**). It has been reported that 5-methylcytosine (5mC) and 5-hydroxymethylcytosine (5hmC), which are precursors in the formation of 5fC (**Fig. 1a**), can stabilize a DNA duplex.[15] In contrast, we observed that

5fC and 5-carboxymethylcytosine (5caC), another product of TET-mediated oxidation, do not stabilize duplexes (**Fig. 2a**).

Circular dichroism (CD) spectroscopy revealed that the 5fC mononucleotide displays an ellipticity maximum of 300 nm, which is redshifted compared to the spectra of other cytosine derivatives (**Supplementary Fig. 4a**). The CD spectrum of the 5fC-containing DNA duplex displayed an absorbance band in the near UV region (λ > 280 nm) as expected, but the ellipticity is negative while for spectra of conventional B-form DNA, ellipticity is positive in this region (**Fig. 2b**). The 5fC DNA spectra is not characteristic of left-handed Z-DNA, because that form presents a negative band in the far UV-region (λ < 200 nm),[16] while 5fC reveals a positive ellipticity in this spectral region. In contrast to the spectra for 5fC, the spectra for 5mC, 5hmC and 5caC are characteristic of B-DNA conformations. These data suggest that 5mC, 5hmC and 5caC do not influence the whole B-DNA double helix structure of CpG repeats containing oligomers, while 5fC drives its conformation to an unusual right-handed helix.

Crystal structure of a formylated CpG repeat

In order to explore the structural consequence of formylated CpG repeats, we then determined the X-ray crystal structure of a self-complementary 5fC-containing dodecamer (5'-CTA**5fC**G**5fC**G**5fC**GTAG-3', ODN6) at 1.40 Å resolution. It is noteworthy that CD spectroscopic analysis of the dodecamer in the crystallization buffer also shows a negative ellipticity in the near UV region (**Supplementary Fig. 4b**), suggesting that the crystal and

the solution structures are conformationally similar. The structure was solved using experimentally derived phases from the anomalous dispersion signal (P-SAD) of the DNA phosphorus atoms (**Table 1**, **Supplementary Fig. 4c**). The refined structure reveals an unusual right-handed helix that is underwound compared to the A-form, and displays 13 bases per turn with altered groove geometry (**Fig. 2c)**. As expected the formyl group of the modified cytosines project into the major groove of the helix. Hydrogen bonds between the formyl group and the exocyclic amino group on C4 lock the rotation of the bond linking the C5 and C(formyl) groups in each 5fC, resulting in a single conformation.

The electron density for the formyl groups of each of the 5fC bases is well-defined and reveals their interactions in detail. The formyl substituent is at the hub of extensive hydration network, and **Fig. 2d** shows the main interactions between water molecules (W1–51), the phosphate backbones, formylcytosines (5fC4, 5fC6 and 5fC8) and adjacent nucleobases (G5, G7 and G9). Each formyl group is networked to the phosphodiester backbone through interactions with four water molecules in the major groove. A hydrogen-bonded water bridges the formyl group of 5fC'8 and O6 of the 3'-adjacent G'9 (W'25). Similarly, a bridging water connects the formyl group of 5fC'8 with O6 of the 5'-adjacent G'7 (W1), and another links the same formyl group to the 3' and 5'-OP1 of the G'7 phosphate backbone (W10). Bases 5fC4 and 5fC'8 are linked through an intricate water bridge comprising formyl–5fC'8–W32–W51–formyl–5fC4. Very similar interactions are observed around the formyl groups of 5fC6 and 5fC4: a water links 5fC6–formyl group with O6 of the 3' adjacent G7, and another bridges 5fC6 to the phosphate backbone

and two others join 5fC6 with 5fC'6. These bridging waters create a secondary network of water molecules lying in the major groove of the helix that are stabilized by the formyl groups of the modified cytosines and the O6 of guanines. Thus, the formyl groups are at the hub of networks that link the phosphate backbone, adjacent nucleobases, and an extensive hydration pattern in the major groove.

Effect of 5fC on the geometry of base pairings

An additional structural consequence of the formyl groups on the cytosines is to affect the geometry of base pairings, and this creates local distortions of the helix. **Fig. 2e** highlights the stacking of the base pairings 5fC4–G'9 and G5–5fC'8. Although the canonical Watson-Crick pairing is conserved, interactions involving the formyl cytosines and water molecules create an unusual base pairing geometry. W'25 creates a bridge between the formyl group of 5fC'8 and O6 of G'9, while W25 bridges 5fC4 and G5. These interactions turn the formylcytosines toward the 3'-adjacent base and push the guanines toward the exterior of the helix. As a result local rotational helix parameters are highly affected and are distinct from those observed in B- or A-DNA. Locally at the base pair 5fC4–G'9 we observe a propeller twist of –18.1, a value nearly double that of canonical C–G base pairs in A- and B-DNA, which have angles of –9.2 ± 4.8° (mean values ± *s.d.*; *n* = 24) and –8.8 ± 9.1° (mean values ± *s.d.*; *n* = 20), respectively. Similarly, we observe distinctive opening angle of –3.2° whereas angles of 1.6 ± 3.0° (mean values ± *s.d.*; *n* = 24) and –0.2 ± 2.3° (mean values ± *s.d.*; *n* = 20) are observed at

canonical CG base pairs in A- and B-DNA respectively.

The 5-formylcytosines directly affect the geometry of the stacking of neighboring nucleotides. **Fig. 2f** highlights the stacking of the paired bases G5–5fC'8 and 5fC6–G'7. It is noteworthy that there is an overlap between the π-system of the formyl groups and that of the N7-C8 of guanines. Additionally the internal hydrogen bond between the formyl group and N4 of the modified cytosine confers to the modified cytosine the appearance of a purine, but with an unusual orientation that approximates an *anti*- orientation about the base-glycosidic bond.

The distinctive local rotational helix parameters and purine-like character of the 5-formylcytosine substantially influence the geometry of the helix by altering base-step parameters. Due to high local propeller angle, we observe a periodic pattern with values between 13.5 ± 2.5° and 4.6 ± 0.7° (mean values ± *s.d.*; *n* = 5) for the roll angle of 5fC–G/G–5fC and G–5fC/G–5fC base steps respectively (**Fig. 3a**). In contrast, no obvious correlations are observed in canonical A- and B-DNA. Furthermore, the altered base pair stacking (**Fig. 2e** and **2f**) influences directly the base-step parameters such as the shift displacement (**Fig. 3b, Supplementary Fig. 5**). We observed a local inversion in translocational parameters at the central 5fC–G/G–5fC step. Around this inflexion point we observe shift values of 0.5 Å, which are among the highest values observed in canonical A- and B-DNA for C–G/G–C and G–C/C–G steps (0.1 ± 0.6 Å (mean values ± *s.d.*; *n* = 14)). It is noteworthy that similar alterations can be observed on hemiformylated 5fC–G/G–C steps. The 1.60 Å resolution crystal structure of a formylated Dickerson-Drew duplex (Kimura *et al.*, unpublished work, PDB entry 1VE8) reveals unusual twist and

roll angles at hemiformylated 5fC sites (**Supplementary Fig. 6**). The hydration network observed within the structure reported in the current work stabilizes the specific conformation of the 5fC–G/G–5fC steps.

Effect of 5fC on helical coiling and trajectory

The unusual local rotational and translocational parameters of the 5fC–G/G–5fC steps impact on DNA helical coiling and trajectory. **Fig. 3c** and **3d** show the geometry of the major and minor groove and emphasize the differences between formylated DNA and canonical A- and B-DNA. Formylation of the CpG repeats narrow the major groove of the helix (**Fig. 3c**) while they open the minor groove (**Fig. 3d**). Interestingly, in the center of the helical axis the minor groove has nearly no depth. This observation reflects the shift of the base pairs toward the exterior of the helix. Due to the distinctive spectroscopic and structural properties of 5fC containing double-stranded DNA, we propose to designate it as F-DNA.

We assessed the effect of incorporation of formylcytosines on longer duplexes by modeling the junctions between the determined structure and a standard model of B-DNA. Using calculated base pair parameters, we generated a 36-mer with B-DNA geometry and another from the solved 5fC-containing 12-mer with flanking ideal B-form helices (**Fig. 3e)**. While the B-DNA presents a uniform structure, the mixed model clearly shows that the introduction of 5fCpG repeats alters the helical trajectory. This gives rise to marked local variation of the grooves creating potential protein recognition sites in the minor groove while displaying a deep binding pocket in the major

groove.

F- to B-DNA conformational transition upon 5fC removal

We studied the dynamic changes to the F-DNA structure upon chemical transformation of 5fC. We monitored by CD spectroscopy the effect of the quantitative $NaBH_4$-mediated reduction of 5fC into 5hmC (**Fig. 4a** and **4b**).[10,17] The spectral data indicate that reduction of 5fC induced a conformational change from F-DNA to B-form DNA. The sigmoidal kinetic profile (**Fig. 4c**) suggests a cooperative structural transition.

Formylation of long oligomers sustain F-DNA formation

CD spectroscopic analysis of C, 5mC, 5hmC and 5fC-containing 147mer DNA duplexes showed that the distinct structural characteristics of F-DNA are maintained in the context of longer DNA oligomers (ODN7–10, **Fig. 4d**). Titration of increasing concentrations of spermine (**Fig. 4e**), a known condensation agent of nucleic acid structures, led to a tightening of the B-DNA structures, while a structural conversion from F- to B-DNA was observed for 5fC-containing oligomers. This result suggests a dynamic equilibrium between F- and B-DNA, and our crystal structure suggests that the equilibrium is likely to be modulated by the hydration of the grooves of the duplex.

We also assessed the impact of 5fC-density on DNA structure. Five different oligomers with varying 5fC density (from 2% to 18% of total base

composition) were analyzed by CD spectroscopy (ODN10–14, **Fig. 4f**). The oligomers displaying high densities of 5fC showed negative ellipticities in the near UV region characteristic of F-DNA. With decreasing 5fC density a gradual inversion of the ellipticity was observed. These observations suggest a mechanism for the interconversion of two well-defined DNA structures that depends on the addition or removal of 5fC.

DISCUSSION

The 5fC containing duplex structure reported here provides new insights into how chemical modifications can affect the structure of DNA at the molecular level. By studying a biologically relevant sequence context we showed that formylation of CpG repeats confers a change in the physical properties of the DNA double-stranded helix. While 5fC did not affect the thermodynamic stability of unmodified CpG repeats containing oligomers, our results demonstrated its ability to drive their structures to a distinct conformation, F-DNA, characterized by helical under-winding. Formylation of CpG repeats is then expected to affect local DNA supercoiling and packaging in chromatin. The enrichment of highly formylated CpG repeats in introns of genes suggest that TET-mediated formylation of genomic DNA may contribute to the control of gene expression by modifying the physical properties of DNA.

Recent proteomics experiments using probes comprising a high density of formylated CpGs, with the propensity to form F-DNA, revealed that 5fC can recruit specific proteins that include glycosylases, transcription regulators and

chromatin remodelers.[5,18] We propose that F-DNA may directly control the recruitment of 5fC readers at formylated sites of the genome. The recognition of DNA structure, rather than the modified bases *per se*, might trigger biological events. The observed alteration of the geometry of the DNA double helix grooves creates potential protein recognition sites. It is noteworthy that discrimination between the different 5-substituents of cytosine by glycosylases using a base-flipping mechanism, for example, does not occur through the creation of protein side-chain interactions in the major groove but rather by probing the minor groove of the DNA substrate.[4,19] Mutational analysis of the catalytic domain of human TDG shows that the P–G–S loop interacting with the major groove of the DNA substrate in the post-reactive complex is unlikely to play a role in discriminating between the different modified cytosines.[4] The base excision repair glycosylase, MPG (also known as AAG), which shows selectivity for 5fC-containing oligonucleotides over other modified cytosines,[5,18] uses a mechanism in which base flipping is initiated through minor groove invasion without any interaction in the major groove.[19] Therefore the opening of the minor groove induced by F-DNA formation could have an impact on 5fC-mediated biological function.

While the structure reported here provides only a static snapshot of the possible conformational diversity of F-DNA, structural analysis of longer (>100 base pairs) double-stranded oligomers bearing different densities of 5fC showed that 5fC alters the classical B-DNA conformation. Furthermore, we have shown that chemical reduction of 5fC to 5hmC induced a conformational change into B-DNA, which highlights the dynamic property of DNA structure upon chemical modification triggered *in vivo* by the TET and TDG enzymes.

We anticipate that further investigations will reveal the full impact of F-DNA on mammalian (and other) genomes.

ACCESSION CODES

The atomic coordinates and structure factors of the reported crystal structure of 5fC oligonucleotide have been deposited to the Protein Data Bank (PDB) under accession code 4QKK.

ACKNOWLEDGEMENT

E. A. Raiber is a Herchel Smith Fellow. The Balasubramanian lab is supported by a Senior Investigator Award from the Wellcome Trust (099232/Z/12/Z to S.B.). The S. Balasubramanian lab also receives core funding from Cancer Research UK (C9681/A11961 to S.B.). D. Y. Chirgadze is supported by the Crystallographic X-ray Facility (CXF) at the Department of Biochemistry, University of Cambridge and B. F. Luisi by the Wellcome Trust (076846/Z/05/A to B.F.L.). We thank the staff of Soleil and Diamond Light Source for use of facilities. We thank Chris Calladine for stimulating discussions.

AUTHOR CONTRIBUTIONS

E. A. Raiber, P. Murat and S. Balasubramanian designed the project and wrote the manuscript with contributions from all authors. E. A. Raiber and P.

Murat performed biophysical experiments and analysed X-ray crystallographic data. D. Y. Chirgadze and B. Luisi acquired and analysed X-ray crystallographic data, D. Y. Chirgadze solved the structure using P-SAD technique. D. Beraldi performed computational analysis of sequence datasets. S. Balasubramanian supervised the project. All authors have interpreted the data, read and approved the manuscript.

REFERENCES

1 Ito, S. *et al.* Tet proteins can convert 5-methylcytosine to 5-formylcytosine and 5-carboxylcytosine. *Science* **333**, 1300-1303 (2011).

2 Globisch, D *et al.* Tissue distribution of 5-hydroxymethylcytosine and search for active demethylation intermediates. *PLoS One* **5**, e15367 (2010).

3 Maiti, A. & Drohat, A. C. Thymine DNA glycosylase can rapidly excise 5-formylcytosine and 5-carboxylcytosine: potential implications for active demethylation of CpG sites. *J. Biol. Chem.* **286**, 35334-35338 (2011).

4 Hashimoto, H., Hong, S., Bhagwat, A. S., Zhang, X. & Cheng, X. 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.* **41**, 10203-10214 (2012).

5 Iurlaro, M. *et al.* A screen for hydroxymethylcytosine and formylcytosine binding proteins suggests functions in transcription and chromatin regulation. *Genome Biol.* **14**, R119 (2013).

6 Renciuk, D., Blacque, O., Vorlickova, M. & Spingler, B. Crystal structures of B-DNA dodecamer containing the epigenetic modifications 5-hydroxymethylcytosine or 5-methylcytosine. *Nucleic Acids Res.* **41**, 9891-9900 (2013).

7 Lercher, L. *et al.* Structural insights into how 5-hydroxymethylation influences transcription factor binding. *Chem. Commun.* **50**, 1794-1796 (2014).

8 Wang, L. *et al.* Programming and Inheritance of Parental DNA Methylomes in Mammals. *Cell* **157**, 979-991 (2014).

9 Raiber, E. A. *et al.* Genome-wide distribution of 5-formylcytosine in embryonic stem cells is associated with transcription and depends on thymine DNA glycosylase. *Genome Biol.* **13**, R69 (2012).

10 Song, C. X. *et al.* Genome-wide profiling of 5-formylcytosine reveals its roles in epigenetic priming. *Cell* **153**, 678-691 (2013).

11 Shen, L. *et al.* Genome-wide analysis reveals TET- and TDG-dependent 5-methylcytosine oxidation dynamics. *Cell* **153**, 692-706 (2013).

12 You, C. *et al.* Effects of Tet-mediated oxidation products of 5-methylcytosine on DNA transcription in vitro and in mammalian cells. *Sci. Rep.* **4:7052** (2014).13

13 Hu, L. *et al.* Crystal Structure of TET2-DNA Complex: Insight into TET-Mediated 5mC Oxidation. *Cell* **155**, 1545-1555 (2013).

14 Xu, L. *et al.* Pyrene-Based Quantitative Detection of the 5-Formylcytosine Loci Symmetry in the CpG Duplex Content during TET-Dependent Demethylation. *Angew. Chem. Int. Ed. Engl.* DOI: 10.1002/ange.201406220 (2014).

15 Thalhammer, A., Hansen, A. S., El-Sagheer, A. H., Brown, T. & Schofield, C. J. Hydroxylation of methylated CpG dinucleotides reverses stabilisation of DNA duplexes by cytosine 5-methylation. *Chem. Commun.* **47**, 5325-5327 (2011).

16 Sutherland, J. C., Griffin, K. P., Keck, P. C. & Takacs, P. Z. Z-DNA: vacuum ultraviolet circular dichroism. *Proc. Natl Acad. Sci. USA* **78**, 4801-4804 (1981).

17 Booth, M. J., Marsico, G., Bachman, M., Beraldi, D. & Balasubramanian, S. Quantitative sequencing of 5-formylcytosine in DNA at single-base resolution. *Nat. Chem.* **6**, 435-440 (2014).

18 Spruijt, C. G. *et al.* Dynamic readers for 5-(hydroxy)methylcytosine and its oxidized derivatives. *Cell* **152**, 1146-1159 (2013).

19 Wyatt, M. D., Allan, J. M., Lau, A. Y., Ellenberger, T. E. & Samson, L. D. 3-Methyladenine DNA glycosylases: structure, function, and biological importance. *BioEssays* **21**, 668–676 (1999).

FIGURE LEGENDS

Figure 1 | High level of cytosine formylation in genomic DNA is observed at CpG repeats. (a) Interplay of modified cytosines in mammalian genomes. Methylation and oxidation of cytosine to 5mC, 5hmC, 5fC and 5caC is mediated by the DNA methyltransferases (DNMT) and the ten-eleven translocation (TET) family of enzymes. The proposed active demethylation pathway relies on the base excision repair (BER) mediated by thymine DNA glycosylase (TDG). **(b)** Influence of the length of CpG repeats, $d(CG)_n$, on the distribution of significant 5fC sites. For increasing length of the CpG repeat, the percentage of 5fC sites above (orange bars) and below (red bars) the formylation median is plotted (formylation median: 36.0%, where FDR < 0.1). **(c)** Formylation of d(CGCGCG) motifs is uniform within a strand and across both strands. d(CGCGCG) motifs where separated in low (< 33%), medium (between 33 and 66%) and high formylation (≥ 66%) level according to the

5fC highest percentage level in the repeat. The boxplots show the overall formylation of each of the three CpGs. See also **Supplementary Fig. 1d** where CpGs are ordered according to the FDR level instead of by genomic position. **(d)** CpG repeats enriched in 5fC were identified in several genes, including the chromatin remodelers Hdac9 and Usp22 (more examples are reported in **Supplementary Fig. 2**). %5fC denotes the percentage of modified cytosines at a specific location averaged across the cell population. Formylation levels are extracted from the quantitative sequencing at single-base resolution of 5fC in mouse two-cell embryos.[8]

Figure 2 | 5fC-containing oligonucleotides are characterized by unusual spectroscopic and structural signatures. (a) UV melting studies showed that the presence of 5fC (green) and 5caC (brown) in a decamer results in a Tm similar to that of cytosine DNA (grey), whereas 5mC (blue) and 5hmC (red) induce stabilisation (data represent mean values ± *s.d.*; *n* = 3). **(b)** CD analysis of cytosine (grey), 5mC (blue), 5hmC (red), 5fC (green) and 5caC (brown)-containing oligonucleotides in comparison to Z-DNA (dotted) revealed a distinct spectroscopic signature associated with 5fC-containing oligonucleotides (y–axis units are not applicable for the Z-DNA sample (poly(dG–dC), 25 mg.mL^{-1}, 3.4 M $NaOCl_4$)). **(c)** Crystal structure overview (top and side view) of a 5fC-containing dodecamer showed formyl groups (green spheres) pointing towards the major groove. A single strand occupied the asymmetric unit, and the duplex was obtained by the application of crystallographic 2-fold symmetry. **(d)** Hydrogen bonding of the formyl groups of 5fCs. Each formyl group is linked with water molecules and interacts with the phosphate backbone and adjacent nucleotides. A secondary network (red

lines) of water molecules lies in the major groove of the helix and is stabilized by the formyl groups of the modified cytosines and the O6 of guanines. **(e)** Bridging water molecules between the formyl groups and the O6 of guanines supports base pair stacking between 5fC4–G'9 and G5–5fC'8. **(d)** Overlap between the π–systems of 5fC and guanines. The base-stacking geometry results in an unusual twist of the helix.

Figure 3 | Comparison of base-step and groove parameters of the 5fC-containing duplex (F-DNA) with B- and A-form of DNA. (a) Roll and **(b)** shift local rotational and translational base-step parameters of F-DNA. The presence of 5fC locally results **(c)** in narrowing the major groove and **(d)** in opening the minor groove of the helix. F-DNA helix parameters (red line) are compared to canonical A- and B- DNA (blue and grey lines respectively). The presented values are the mean (line, *n* = *3*) and standard deviation (colored area) obtained from experimental structures of A-DNA and B-DNA of similar length and base composition (see **Online Methods**). **(e)** Modeling of a 36-mer with B-DNA geometry and another from the X-ray structure of the 5fC-containing dodecamer with flanking ideal B-form helices DNA showed alteration of the helical trajectory and marked local variation of the grooves induced by 5fC.

Figure 4 | Induced conformational transformation of F- to B-DNA. (a) HPLC traces of digested oligomer before (top panel) and after (bottom panel) chemical reduction using aqueous $NaBH_4$. **(b)** Time-dependent structural conversion of F-DNA to B-DNA upon 5fC $NaBH_4$-mediated reduction as

monitored by CD spectroscopy. CD spectra were acquired every 3 min over a period of 45 min. Shift in the band during $NaBH_4$ reduction (blue to red) indicates a structural change in the DNA conformation, which was confirmed by **(c)** the kinetic profile monitored at 285nm (blue) and 275nm (red). **(d)** The long oligomer (147mer) containing 5fC (green) showed the characteristic inverted band of F-DNA in the near UV region. **(e)** CD spectra of 5hmC (grey) and 5fC (green)-containing long oligomers in the absence (dotted line) or presence (plain lines) of 200mM of spermine. High concentration of spermine resulted in an inversion of the molar ellipticity in the near UV region for F-DNA. **(f)** Decrease in 5fC densities (18%–2% of total base composition) resulted in gradual inversion of the ellipticity in the near UV region.

TABLE

Table 1 Data collection, phasing using phosphorous SAD (P-SAD) and refinement statistics

	Native dataset	Phosphorous SAD dataset
Data collection		
Space group	$P4_32_12$	$P4_32_12$
Cell dimensions		
a, *b*, *c* (Å)	44.39 44.39 46.25	44.64 44.64 45.94
α, β, γ (°)	90.0 90.0 90.0	90.0 90.0 90.0
Resolution (Å)	46.25 – 1.40 (1.48 – 1.40)[a]	45.94 – 1.60 (1.63 – 1.60)
R_{merge}	6.0 (71.5)	7.3 (60.3)
$I / \sigma I$	19.3 (2.0)	48.8 (4.9)
Completeness (%)	100 (100)	100 (100)
Redundancy	12.1 (12.2)	85.1 (40.6)
Refinement		
Resolution (Å)	46.25 – 1.40	
No. reflections	9,608	
R_{work} / R_{free}	0.140 / 0.159	
No. atoms		
DNA	249	
Water	54	
B factors		
DNA (Å^2)	28.1	
Water (Å^2)	46.5	
r.m.s deviations		
Bond lengths (Å)	0.014	
Bond angles (°)	1.377	

Both datasets were collected from a single crystal.

[a]Values in parentheses are for highest-resolution shell.

ONLINE METHODS

Sample Preparation

DNA oligonucleotides (ODN 1–5) were purchased from Eurogentec. ODN1–4 were prepared in phosphate buffer saline (PBS) and annealed by heating to 95 °C for 5 min and cooling to room temperature at a rate of 0.1 °C.sec^{-1}. The Z-DNA structure was obtained by annealing poly(dG-dC) (Sigma) in PBS supplemented with 3.4 M $NaOCl_4$ at a final concentration of 25 mg.mL^{-1}. ODN6–10 were obtained by PCR using the Dreamtaq polymerase (Fermentas) and modified deoxytriphosphates (Trilink). The DNA was subsequently purified using the GeneJet PCR purification kit and eluted in 10mM sodium cacodylate buffer. Refer to **Supplementary table 1** for sequences.

UV spectroscopy

UV melting curves were collected using a Varian Cary 400 Scan UV-visible spectrophotometer by following the absorbance at 260 nm. Oligonucleotides solutions were prepared at final concentrations of 4 µM in PBS. The samples were annealed by heating to 95 °C for 10 min and then slowly cooled to room temperature at a rate of 0.1 °C.sec^{-1}. Each sample was transferred to a quartz cuvette with 1 cm path length, covered with a layer of mineral oil,

placed in the spectrophotometer and equilibrated at 5 °C for 10 min. Samples were then heated to 95 °C and cooled to 5 °C at a rate of 1 °C.min^{-1}, with data collection every 1 °C during both melting and cooling. Melting temperature (Tm) values were obtained from the minimum of the first derivative of the melting curve.

Circular Dichroism spectroscopy

CD spectroscopy experiments were conducted on a Chirascan Plus spectropolarimeter using a quartz cuvette with an optical path length of 1 mm. Oligonucleotide solutions were prepared at a final concentration of 1 to 10 μM in either PBS or 10 mM lithium cacodylate (pH 7.2). The samples were annealed by heating at 95 °C for 10 min and slowly cooled to room temperature at a rate of 0.1 °C.sec^{-1}. Scans were performed over the range of 200–320 nm at 25 °C. Each trace was the result of the average of three scans taken with a step size of 1 nm, a time per point of 1 s and a bandwidth of 1 nm. A blank sample containing only buffer was treated in the same manner and subtracted from the collected data. The data were finally baseline corrected at 320 nm.

Preparation of crystals

ODN5 was dissolved in water, desalted using a PD10 column (GE Healthcare) and annealed by heating to 95 °C for 5 min and cooling to room temperature at a rate of 0.1 °C.sec^{-1}. Crystallization trials were performed by the vapour diffusion sitting-drop technique in 96-well MRC 2-drop crystallization plates (SWISSCI AG) using Nucleix, MPD and PEGS I

crystallisation screens (Qiagen Ltd.). □□□□nL of the crystallisation screen conditions were mixed with 200 nL of 5fC oligonucleotide at the concentrations of 1 mM and 0.1 mM, and set against 70 uL of reservoir using a crystallization robot (Crystal Phoenix, Art Robbins Instruments, Inc.). The crystallization trials were incubated at 19 °C and crystal growth monitored with a Rock Imager 1000 (Formulatrix, Inc.). Several conditions produced crystals, which appeared after 2 days and grew to maximum size (0.5 x 0.3 x 0.3 mm^3) after about 1–2 weeks. The crystals used for X-ray diffraction data collection grew from crystallisation buffer comprised of 0.01 M magnesium sulphate, 0.05 M sodium cacodylate pH 6.0, 1.8 M lithium sulphate.

Diffraction Data Collection and Processing

Crystals were cryoprotected by immersing in crystallization condition with 26% *v/v* ethylene glycol for a few seconds then flash frozen in liquid nitrogen. High redundancy phosphorus single wavelength anomalous dispersion (P-SAD) dataset was collected using a copper rotating anode X-ray diffraction system equipped with confocal mirror monochromator, a kappa geometry goniometer, and Platinum 135 CCD detector (PROTEUM X8, Bruker AXS, Ltd) at 100K using a COBRA Cryostream cryogenic cooling device (Oxford Cryosystems, Ltd). Phosphorus has a weak anomalous scattering signal at the 1.5418 Å wavelength used for data collection (f'' = 0.43 e). However, by collecting highly redundant data, the anomalous signal-to-noise level in the dataset is increased to the point where it can be recorded with sufficient accuracy to successfully determine phases. The dataset was collected using a specific data collection strategy protocol that maximizes the redundancy of data in the

high-resolution shell to about 40 (mean redundancy of the dataset was 85). The resolution of the dataset was manually limited to 1.60 Å. The exposure time was set to 15 sec for a single phi-oscillation image of 1 degree, and the total of 2,505 oscillation images were collected in 31 different kappa geometry orientations. The dataset was indexed, scaled and merged using PROTEUM2.[20] The crystal belongs to tetragonal $P4_32_12$ space group with cell parameters $a = b = 44.6$ Å, c = 45.9 Å, $\alpha = \beta = \gamma = 90°$ and contained one molecule of 5fC oligonucleotide (dodecamer) in the asymmetric portion of the unit cell. A high-resolution native dataset was collected at the Diamond Light Source synchrotron science facility (Oxford, United Kingdom) beamline I24 equipped with Pilatus 6M pixel array detector (DECTRIS, Ltd) the X-ray wavelength was set to 0.9686 Å, the crystal was kept at 100K during data collection. A total of 1,800 phi oscillation images of 0.1 degree at 0.1 seconds exposure were collected. The crystal diffracted to a maximum resolution of 1.40 Å. The diffraction data were indexed, scaled and merged using XDS software.[21] The crystallographic data collection statistics are summarized in **Table 1**.

Crystal Structure Determination, Model Building and Refinement

Experimental phases were obtained from the P-SAD dataset collected from the in-house source. The PHENIX software suite was used for performing of all of the crystallographic calculations for structure solution and refinement.[22] The analysis of anomalous measurability in the P-SAD dataset as defined by PHENIX demonstrated the presence of statistically significant anomalous signal to 2.2 Å resolution. The anomalous atom substructure determination

identified the position of 11 out of 11 possible phosphorus sites in the asymmetric unit. Phases were calculated using Phaser (Figure of Merit 0.54) and further improved by electron density modification using RESOLVE (Figure of Merit 0.74). The resulting experimental electron density map was readily interpretable (**Supplementary Fig. 4c**), and an initial model built using molecular graphics software suite COOT.[23] The initial model of 5fC oligonucleotide was refined against high-resolution native dataset at 1.40 Å which had been collected at Diamond Light Source synchrotron facility (beamline I24). Solvent molecules were added manually and through an automated procedure as implemented in the PHENIX refinement protocols. All B-factors of the DNA molecule were refinement anisotropically. Hydrogen atoms were added in their riding positions to the DNA atoms but not to the water molecules. The R_{cryst} and R_{free} converged to the values of 14.0% and 15.9%, respectively. The crystallographic statistics and structural validation details are shown in **Table 1**.

Structure Analysis

Helix, base and base pair parameters were calculated with 3DNA or curve+ software packages.[24,25] The values for A- and B-DNA were obtained from experimental structures of A-DNA (PDB-IDs: 117D, 116D and 1QPH)[26.27] and B-DNA (PDB-IDs: 1BNA, 1HQ7 and 119D).[28–30]

Chemical conversion

ODN5 was annealed in PBS at a concentration of 10 μM and subjected to CD analysis in a quartz cuvette with a path length of 0.1 cm. At t = 0, a freshly prepared aqueous $NaBH_4$ solution (1M) was added directly in the cuvette at a final concentration of 10 mM. CD spectra were acquired every 3 min for 45 min. The cuvette was regularly shaken to avoid formation of bubbles that disturb collection of CD spectra. The reaction was quenched by the addition of an equal volume of methanol. The sample was subsequently used for DNA digestion and HPLC analysis.

DNA digestion and HPLC analysis

Oligonucleotides were digested using the DNA Degradase Plus (Zymo Research), purified with Amicon Ultra 0.5 mL 10 kDa columns and analysed by HPLC using an Agilent 1100 HPLC with a flow of 1 $mL.min^{-1}$ over an Eclipse XDB-C18 3.5 μm, 3.0 x 150 mm column. The column temperature was maintained at 45 °C. Eluting buffers were buffer A (500 mM Ammonium Acetate (Fisher) pH 5), Buffer B (Acetonitrile) and Buffer C (Water). Buffer A was held at 1 % throughout the whole run and the gradient for the remaining buffers was 0 min – 0.5 % B, 2 min – 1 % B, 8 min – 4 % B, 10 min – 95 % B.

METHODS-ONLY REFERENCES

20 PROTEUM 2 User Manual, Bruker AXS, 2010.

21 Kabsch, W. Integration, scaling, space-group assignment and post-refinement. *Acta Crystallogr. D Biol. Crystallogr.* **66**, 133-144 (2010).

22 Adams, P. D. *et al.* PHENIX: a comprehensive Python-based system for macromolecular structure solution. *Acta Crystallogr. D Biol. Crystallogr.* **66**, 213-221 (2010).

23 Emsley, P., Lohkamp, B., Scott, W.G. & Cowtan, K. Features and development of Coot. *Acta Crystallogr. D Biol. Crystallogr.* **66**, 486-501 (2010).

24 Zheng, G., Lu, X.J. & Olson, W.K. Web 3DNA—a web server for the analysis, reconstruction, and visualization of three-dimensional nucleic-acid structures. *Nucleic Acids Res.* **37** (Web Server issue), W240–W246 (2009).

25 Lavery, R., Moakher, M., Maddocks, J.H., Petkeviciute, D. & Zakrzewska, K. CURVES+ web server for analyzing and visualizing the helical, backbone and groove parameters of nucleic acid structures. *Nucleic Acids Res.* **37**, 5917-5929 (2009).

26 Bingman, C., Jain, S., Zon, S. & Sundaralingam, M. Crystal and molecular structure of the alternating dodecamer d(GCGTACGTACGC) in the A-DNA form: comparison with the isomorphous non-alternating dodecamer d(CCGTACGTACGG). *Nucleic Acids Res.* **20**, 6637-6647 (1992).

27 Bingman, C. A., Zon, G. & Sundaralingam, M. Crystal and molecular structure of the A-DNA dodecamer d(CCGTACGTACGG). Choice of fragment helical axis. *J. Mol. Biol.* **227**, 738-756 (1992).

28 Drew, H. R. *et al.* Structure of a B-DNA dodecamer: conformation and dynamics. *Proc. Natl Acad. Sci. USA* **78**, 2179-2183 (1981).

29 Locasale, J. W., Napoli, A.A., Chen, S., Berman, H.M. & Lawson, C.L. Signatures of protein-DNA recognition in free DNA binding sites. *J. Mol. Biol.* **386**, 1054-1065 (2009).

30 Leonard, G. A. & Hunter, W. N. Crystal and molecular structure of d(CGTAGATCTACG) at 2.25 A resolution. *J. Mol. Biol.* **234**, 198-208 (1993).