5-Formylcytosine alters the structure of the DNA double helix

Eun-Ang Raiber1,5, Pierre Murat1,5, Dimitri Y Chirgadze2, Dario Beraldì3, Ben F Luisi2 & Shankar Balasubramanian1,3,4

The modified base 5-formylcytosine (5fC) was recently identified in mammalian DNA and might be considered to be the ‘seventh’ base of the genome. This nucleotide has been implicated in active demethylation mediated by the base excision repair enzyme thymine DNA glycosylase. Genomics and proteomics studies have suggested an additional role for 5fC in transcription regulation through chromatin remodeling. Here we propose that 5fC might affect these processes through its effect on DNA conformation. Biophysical and structural analysis revealed that 5fC alters the structure of the DNA double helix and leads to a conformation unique among known DNA structures including those comprising other cytosome modifications. The 1.4-Å-resolution X-ray crystal structure of a DNA dodecamer comprising three 5fCpG sites shows how 5fC changes the geometry of the grooves and base pairs associated with the modified base, leading to helical underwinding.

To date, four modified cytosines have been discovered in mammalian genomes: 5-methylcytosine (5mC), 5-hydroxymethylcytosine (5hmC), 5fC and 5-carboxycytosine (5caC). The discovery of these naturally occurring nucleobases has sparked the search for possible associated biological functions1,2. The most frequently postulated function is their role in the active DNA demethylation pathway (Fig. 1a), a key process in resetting epigenetic information. A vital player in this pathway is the thymine DNA glycosylase (TDG), which can excise both 5fC and 5caC but prefers the former3. However, the mechanism by which TDG recognizes the oxidized products remains unclear4. Recently the identification of transcriptional regulators, DNA-repair factors and chromatin regulators that selectively bind to 5fC in genomic sequences has suggested that 5fC may be an epigenetic signal in its own right5.

Because 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 structural properties of the DNA double helix. Previous reports have shown that 5mC and 5hmC do not influence either the B-DNA double-helix structure or the geometry of the modified base pair6,7. 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 in 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 affect biology.

RESULTS
Highly formylated elements are prevalent in CpG repeats
Quantitative sequencing of 5fC at single-base resolution, as previously described in mouse embryonic stem cells10 and two-cell embryos8, revealed high levels of formylated cytosine in specific genomic locations. Data extracted from 5fC sequencing of mouse two-cell embryos indicated that the highly formylated elements are found in CpG repeats (d(CG)n, n ≥ 3) (Fig. 1b and 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 and Supplementary Fig. 1d,e), results 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 ten-eleven translocation (TET) enzymes preferentially oxidize 5mC in symmetric methylated CpG sites11 and maintain symmetry of the resulting formylated CpG sites14. Long CpG repeats with high formylation levels (up to 80%) can be observed in genes such as chromatin remodelers (for example, Hdac9 and Usp22) and transcription factors (for example, Maz and Ebf3) (Fig. 1d and 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,c). When 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.

1Department of Chemistry, University of Cambridge, Cambridge, UK. 2Department of Biochemistry, University of Cambridge, Cambridge, UK. 3Cancer Research UK Cambridge Institute, Li Ka Shing Centre, Cambridge, UK. 4School of Clinical Medicine, University of Cambridge, Cambridge, UK. 5These authors contributed equally to this work. Correspondence should be addressed to S.B. (sb10031@cam.ac.uk).

Received 27 October; accepted 21 November; published online 15 December 2014; doi:10.1038/nsmb.2936
Thermodynamic and spectroscopic properties of CpG repeats

In order to assess the impact of cytosine formylation within CpG repeats on the stability and structure of DNA and to compare the effects of 5fC to those of 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,b and Supplementary Table 1). It has been reported that 5mC and 5hmC, which are precursors in the formation of 5fC (Fig. 1a), can stabilize a DNA duplex15. In contrast, we observed that 5fC and 5caC, another product of TET-mediated oxidation, do not stabilize duplexes (Fig. 2a).

CD spectroscopy revealed that the 5fC mononucleotide displays an ellipticity maximum of 300 nm, which is redshifted in comparison 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 was negative, whereas for spectra of conventional B-form DNA, ellipticity is positive in this region (Fig. 2b). The 5fC DNA spectra are not characteristic of left-handed Z-DNA because that form presents a negative band in the far UV-region (λ <200 nm)16, whereas 5fC spectra showed 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 repeat–containing oligomers, whereas 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 1.40Å resolution. It is noteworthy that CD spectroscopic analysis of the dodecamer in the crystallization buffer also showed a negative ellipticity in the near-UV region (Supplementary Fig. 4b), thus suggesting that the crystal and the solution structures are conformationally similar. We solved the structure by using experimentally derived phases from the single-wavelength anomalous dispersion signal of the DNA phosphorus atoms (P-SAD; Table 1 and Supplementary Fig. 4c). The refined structure has an unusual right-handed helix that is underwound compared to the A form and that displays 13 bases per turn with altered groove geometry (Fig. 2c). As expected, the formyl groups 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, thus resulting in a single conformation.

The electron density for the formyl groups of each of the 5fC bases is well defined, showing their interactions in detail. The formyl substituent is at the hub of an extensive hydration network, and interactions between water molecules (W1–W51), the phosphate backbones, formylcytosines (5fC4, 5fC6 and 5fC8) and adjacent nucleobases (G5, G7 and G9) are visible (Fig. 2d). 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 5fC8 and O6 of the 3′-adjacent G’9 (W’25).
Similarly, a bridging water connects the formyl group of 5fC8 with O6 of the 5′-adjacent G7 (W1), and another links the same formyl group to the 3′- and 5′-OP1 of the G7 phosphate backbone (W10). Bases 5fC4 and 5fC8 are linked through an intricate water bridge, formyl-5fC8-W32-W51-formyl-5fC4. We observed very similar interactions around the formyl groups of 5fC6 and 5fC4: one water links the 5fC6 formyl group with O6 of the 3′-adjacent G7, another water bridges 5fC6 to the phosphate backbone and two other waters join 5fC6 with 5fC4. 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 to create local distortions of the helix (stacking of the base-pairings 5fC4-G9 and G5-5fC8 in Fig. 2e). Although the canonical Watson-Crick pairing is conserved, interactions involving the formyl cytosines and water molecules create an unusual base-pairing geometry. W25 creates a bridge between the formyl group of 5fC8 and O6 of G9, and 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-G9 we observed 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 ± s.d.; n = 24 base pairs) and −8.8 ± 9.1° (mean ± s.d.; n = 20 base pairs), respectively. Similarly, we observed a distinctive opening angle of −3.2°, whereas we observed angles of 1.6 ± 3.0° (mean ± s.d.; n = 24 base pairs) and −0.2 ± 2.3° (mean ± s.d.; n = 20 base pairs) at canonical CG base pairs in A- and B-DNA, respectively.

The 5-formylcytosines directly affect the geometry of the stacking of neighboring nucleotides (stacking of the paired bases G5-5fC8 and 5fC6-G7 in Fig. 2f). 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 the purine-like character of the 5-formylcytosine substantially influence the geometry of the helix by altering base-step parameters. Because of the high local propeller angle, we observed a periodic pattern with values between
Formylation of long oligomers sustains F-DNA formation

CD spectroscopic analysis of C-, 5mC-, 5hmC- and 5fC-containing 147-mer DNA duplexes showed that the distinct structural
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 found that modification of CpG repeats confers a change in the physical properties of the DNA double-stranded helix. Although 5fC did not affect the thermodynamic stability of unmodified CpG repeat–containing oligomers, our results demonstrated its ability to drive their structures to a distinct conformation, F-DNA, characterized by helical underwinding. Formylation of CpG repeats therefore should affect local DNA supercoiling and packaging in chromatin. The enrichment of highly formylated CpG repeats in introns of genes suggests 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 with a high density of formylated CpGs, with the propensity to form F-DNA, have shown that 5fC can recruit specific proteins that include glycosylases, transcription regulators and chromatin remodelers. We propose that F-DNA may directly control the recruitment of 5fC readers at formylated sites of the genome. The recognition of the altered DNA conformation, 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-position substituents of cytosine by glycosylases using a base-flipping mechanism, for example, occurs not through creating interactions with protein side chains in the major groove but rather by probing the minor groove of the DNA substrate. Mutational analysis of the catalytic domain of human TDG has shown that the P-G-S loop interacting with the major groove of the DNA substrate in the postreactive complex is unlikely to play a part in discriminating between the different modified cytosines. The base excision repair glycosylase MPG (also known as AAG), which shows selectivity for 5fC-containing oligonucleotides over other modified cytosines, uses a mechanism in which base-flipping is initiated through minor-groove invasion without any interaction in the major groove. Therefore the opening of the minor groove induced by F-DNA formation could have an impact on 5fC-mediated biological function.

Although 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 induces a conformational change into B-DNA, highlighting 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.

METHODS

Methods and any associated references are available in the online version of the paper.

Accession codes. Coordinates and structure factors of the crystal structure of 5fC oligonucleotide have been deposited in the Protein Data Bank under accession code 4QKK.
Note: Any Supplementary Information and Source Data files are available in the online version of the paper.

ACKNOWLEDGMENTS

E.-A.R. is supported as a Herchel Smith Fellow. The Balasubramanian laboratory is supported by a Senior Investigator Award from the Wellcome Trust (099232/Z/12/Z to S.B.), and it also receives core funding from Cancer Research UK (C5681/A11961 to S.B.). D.Y.C. is supported by the Crystallographic X-ray Facility (CXF) at the Department of Biochemistry, University of Cambridge, and B.E.L. is supported for the Wellcome Trust (076846/Z/05/A to B.E.L.). We thank the staff of Soleil and Diamond Light Source for use of facilities. We thank C. Calladine for stimulating discussions.

AUTHOR CONTRIBUTIONS

E.-A.R., P.M. and S.B. designed the project and wrote the manuscript with contributions from all authors. E.-A.R. and P.M. performed biophysical experiments and analyzed X-ray crystallographic data. D.Y.C. and B.E.L. acquired and analyzed X-ray crystallographic data. D.Y.C. solved the structure with P-SAD. D.B. performed computational analysis of sequence data sets. S.B. supervised the project. All authors interpreted the data and read and approved the manuscript.

COMPETING FINANCIAL INTERESTS

The authors declare no competing financial interests.

Reprints and permissions information is available online at http://www.nature.com/reprints/index.html.

1. Ito, S. et al. Tet proteins can convert 5-methylcytosine to 5-formylcytosine and 5-carboxylcytosine. Science 333, 1300–1303 (2011).
2. Pfaffeneder, T. et al. The discovery of 5-formylcytosine in embryonic stem cell DNA. Angew. Chem. Int. Ed. Engl. 50, 7008–7012 (2011).
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. 40, 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. & Springer, 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-hydroxymethylcytosine influences transcription factor binding. Chem. Commun. (Camb.) 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. 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. 53, 11223–11227 (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. (Camb.) 47, 5325–5327 (2011).
16. Sutherland, J.C., Griffin, J.C., 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).
ONLINE METHODS
Sample preparation. DNA oligonucleotides (ODN1–5) were purchased from Eurogentec. ODN1–4 were prepared in phosphate-buffered saline (PBS) and annealed by heating to 95 °C for 5 min and cooling to room temperature at a rate of 0.1 °C s−1. The Z-DNA structure was obtained by annealing poly(dG-dC) (Sigma) in PBS supplemented with 3.4 M NaClO4 at a final concentration of 25 mg mL−1. ODN6–10 were obtained by PCR with the DreamTaq polymerase (Fermentas) and modified deoxytriposphates (Trilink). The DNA was subsequently purified with the GeneJet PCR purification kit and eluted in 10 mM sodium cacodylate buffer. Sequences are shown in Supplementary Table 1.

Bioinformatic analysis. Methods and codes used for the analysis of the 5fC quantitative sequencing at single-base-resolution datasets (reported in Fig. 1b–d and Supplementary Figs. 1a–e, 2 and 3) are described in Supplementary Note.

UV spectroscopy. UV melting curves (absorbance at 260 nm) were collected with a Varian Cary 400 Scan UV-visible spectrophotometer. Oligonucleotide solutions were prepared at final concentrations of 4 µM in PBS. The samples were annealed by heating to 95 °C for 10 min and were then slowly cooled to room temperature at a rate of 0.1 °C s−1. Each sample was transferred to a quartz cuvette with a 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 with 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 were slowly cooled to room temperature at a rate of 0.1 °C s−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 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 with 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 vapor-diffusion sitting-drop technique in 96-well MRC 2-drop crystallization plates (Swissci AG) with Nucleix, MPD and PEGS I crystallization screens (Qiagen). The crystallization-screen conditions (200 mL) were mixed with 200 mL of 5fC oligonucleotide at the concentrations of 1 mM and 0.1 mM and were set against 70 µL of reservoir with a crystallization robot (Crystal Phoenix, Art Robbins Instruments). The crystallization trials were incubated at 19 °C, and crystal growth was monitored with a Rock Imager 1000 (Formulatrix). Several conditions produced crystals, which appeared after 2 d and grew to maximum size (0.5 × 0.3 × 0.3 mm3) after about 1–2 weeks. The crystals used for X-ray diffraction data collection grew from crystallization buffer of 0.01 M magnesium sulfate, 0.05 M sodium cacodylate, pH 6.0, and 1.8 M lithium sulfate.

Diffraction data collection and processing. Crystals were cryoprotected by immersion in crystallization condition with 26% v/v ethylene glycol for a few seconds and then were flash frozen in liquid nitrogen. The high-redundancy phosphorus single-wavelength anomalous dispersion (P-SAD) data set was collected with a copper rotating anode X-ray diffraction system equipped with confocal mirror monochromator, a κ-geometry goniometer and a Platinum 135 CCD-detector (PROTEUM X8, Bruker AXS) at 100 K with a Cobra Cryostream cryogenic cooling device (Oxford Cryosystems). Phosphorus has a weak anomalous scattering signal at the 1.5418-Å wavelength used for data collection (fσ = 0.43e). However, by collecting highly redundant data, the anomalous signal-to-noise level in the data set is increased to the point that it can be recorded with sufficient accuracy to successfully determine phases. The data set was collected with a specific data collection–strategy protocol that maximizes the redundancy of data in the highest-resolution shell to about 40 (with a mean redundancy of the data set of 85). The resolution of the data set was manually limited to 1.60 Å. The exposure time was set to 15 s for a single Ψ-oscillation image of 1°, and the total of 2,505 oscillation images were collected in 31 different κ-geometry orientations. The data set was indexed, scaled and merged with PROTEUM2 (Bruker AXS). The crystal belongs to the tetragonal P4321 space group with cell parameters a = b = 44.6 Å, c = 45.9 Å, and α = β = γ = 90°, and it contained one molecule of 5fC oligonucleotide (dodecamer) in the asymmetric portion of the unit cell. A high-resolution native data set was collected at the Diamond Light Source synchrotron science facility (Oxford) beamline I24 equipped with a Pilatus 6M pixel array detector (Dectris). The X-ray wavelength was set to 0.9868 Å, and the crystal was kept at 100 K during data collection. A total of 1,800 Ψ-oscillation images of 0.1° at 0.1-s exposure were collected. The crystal diffraction to a maximum resolution of 1.40 Å. The diffraction data were indexed, scaled and merged with XDHS20. 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 data set collected from the in-house source. The PHENIX software suite was used for all of the crystallographic calculations for structure solution and refinement21. The analysis of anomalous measurability in the P-SAD data set as defined by PHENIX demonstrated the presence of statistically significant anomalous signal to 2.2-Å resolution. The anomalous atom substructure determination identified the positions of 11 out of 11 possible phosphorus sites in the asymmetric unit. Phases were calculated with Phaser (figure of merit 0.54) and further improved by electron density modification with RESOLVE (figure of merit 0.74). The resulting experimental electron density map was readily interpretable (Supplementary Fig. 4c), and an initial model was built with the molecular graphics software suite COOT22. The initial model of 5fC oligonucleotide was refined against a high-resolution native data set at 1.40 Å, which had been collected at the 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 refined anisotropically. Hydrogen atoms were added in their riding positions to the DNA atoms but not to the water molecules. The Rcryst and Rfree 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 packages23–24. The values for A- and B-DNA were obtained from experimental structures of A-DNA (PDB 117D, 116D and 1QP1)25,26 and B-DNA (PDB 18NA, 1HQ7 and 1N9D)27–29.

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 NaBH4 solution (1 M) 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 with the DNA Degradasde Plus (Zymo Research), purified with Amicon Ultra 0.5-mL 10-kdA columns and analyzed by HPLC with an Agilent 1100 HPLC with a flow of 1 mL min−1 over an Eclipse XDB-C18 3.5-μm, 3.0 × 150 mm column. The 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.

20. Kabsch, W. Integration, scaling, space-group assignment and post-refinement. Acta Crystallogr. D Biol. Crystallogr. 66, 133–144 (2010).
21. Adams, P.D. et al. PHENIX: a comprehensive Python-based system for macromolecular structure solution. Acta Crystallogr. D Biol. Crystallogr. 66, 213–221 (2010).
22. Emsley, P., Lohkamo, B., Scott, W.G. & Cowtan, K. Features and development of Coot. Acta Crystallogr. D Biol. Crystallogr. 66, 486–501 (2010).
23. 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, W240–W246 (2009).

24. 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).

25. 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).

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

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

28. 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).

29. 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).