5-Formylcytosine organizes nucleosomes and forms Schiff base interactions with histones in mouse embryonic stem cells

Eun-Ang Raiber, [1]

Guillem Portella, [1]

Sergio Martínez Cuesta, [1,2]

Robyn Hardisty, [1]

Pierre Murat, [1]

Zhe Li, [1]

Mario Iurlaro, [3,8]

Wendy Dean, [3]

Julia Spindel, [3]

Dario Beraldi, [4]

Zheng Liu, [1]

Mark A. Dawson, [5]

Wolf Reik, [3,6]

Shankar Balasubramanian, [1✉,2,7]

Email sb10031@cam.ac.uk

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

[2] Cancer Research UK Cambridge Institute, University of Cambridge, Cambridge, UK

[3] Epigenetics Programme, The Babraham Institute, Cambridge, UK

[4] Cancer Research Centre, Institute of Cancer Sciences, University of Glasgow, Glasgow, UK

[5] Cancer Research Division, Peter MacCallum Cancer Centre, University of Melbourne, Melbourne, Australia

[6] The Wellcome Trust Sanger Institute, Cambridge, UK

[7] School of Clinical Medicine, University of Cambridge, Cambridge, UK

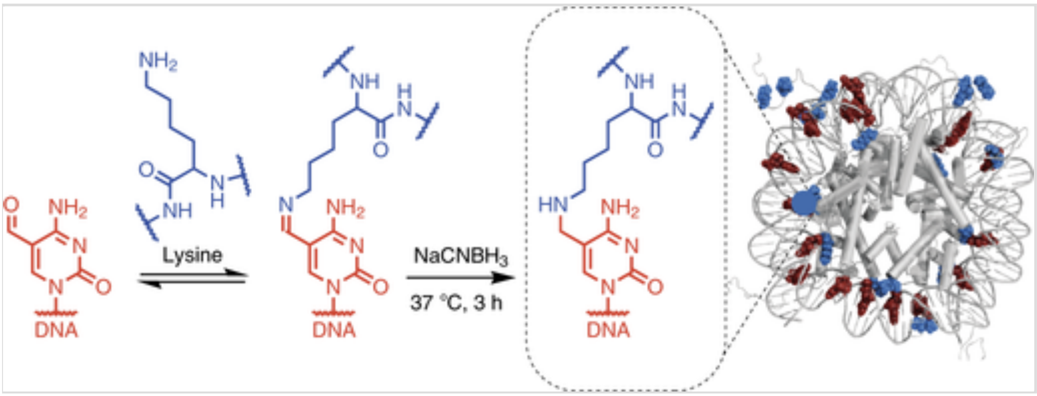
[8] Present Address: Friedrich Miescher Institute for Biomedical Research, Basel, Switzerland

Received: 30 May 2018 / Accepted: 29 August 2018

Abstract

Nucleosomes are the basic unit of chromatin that help the packaging of genetic material while controlling access to the genetic information. The underlying DNA sequence, together with transcription-associated proteins and chromatin remodelling complexes, are important factors that influence the organization of nucleosomes. Here, we show that the naturally occurring DNA modification, 5-formylcytosine (5fC) is linked to tissue-specific nucleosome organization. Our study reveals that 5fC is associated with increased nucleosome occupancy in vitro and in vivo. We demonstrate that 5fC-associated nucleosomes at enhancers in mammalian hindbrain and heart are linked to elevated gene expression. Our study also reveals the formation of a reversible-covalent Schiff base linkage between lysines of histone proteins and 5fC within nucleosomes in a cellular environment. We define their specific genomic loci in mouse embryonic stem cells and look into the biological consequences of these DNA–histone Schiff base sites. Collectively, our findings show that 5fC is a determinant of nucleosome organization and plays a role in establishing distinct regulatory regions that control transcription.

A series of in vitro and in vivo studies has now shown that 5fC is linked to increased nucleosome occupancy and stability. Moreover, there is evidence that Schiff base formation between histones and 5fC impacts RNA polymerase II transcription activity in mouse embryonic stem cells.



These authors contributed equally: Eun-Ang Raiber, Guillem Portella.

Main

The organization of DNA by nucleosomes is an important feature of chromatin structure. While the DNA sequence is known to be a determinant of where nucleosomes form[1, 2, 3], in vitro studies have shown that DNA base modifications have the potential to alter the conformation and stability of nucleosomes[4, 5, 6, 7]. Other studies in mouse embryonic stem cells (mESCs) have linked the presence of 5-hydroxymethylcytosine (5hmC) to the depletion of nucleosome occupancy, a correlation that was not observed for 5-formylcytosine (5fC)[8]. 5fC is a natural, modified base that is generated by the oxidation of 5-methylcytosine (5mC) by ten eleven translocation (TET) enzymes[9]. The 5fC base can be removed via base excision repair catalysed by thymine DNA glycosylase (TDG)[10] and it has been proposed to mark sites that undergo active demethylation[11]. However, we recently demonstrated that the majority of 5fC sites are stable and persist in genomic DNA in vivo[12], with a genome-wide distribution that is tissue-dependent in mice[13]. Moreover, the nature of the formyl group confers specific chemical properties to 5fC that are distinct from other cytosine modifications such as 5hmC and 5-carboxycytosine (5caC). Indeed, recent studies using proteomics[14] and gel shift analysis[15] have shown that 5fC can form a Schiff base conjugate with histone proteins in vitro. The demonstration of a 5fC–histone interaction within a cellular, chromatin context remains elusive, as is an understanding of the biological significance of this reversible, covalent crosslink. It has also been demonstrated that just a single 5fC unit is sufficient to increase the flexibility of the DNA strand[16]. Since the sequence-dependent mechanical plasticity required for the DNA to bend and adopt the nucleosome structure is central for nucleosome organization, physical changes to the DNA at 5fC sites might locally impact chromatin structure. Here, we compare the effects of various cytosine DNA modifications on nucleosome occupancy and stability within an unnatural and genomic DNA sequence context. We elucidate the role of 5fC on nucleosome organization and gene regulation in mouse embryonic tissues and also demonstrate the formation of a covalent linkage between 5fC and histone protein H3 in mESCs. Our data support the existence of a molecular linkage between DNA base modification and chromatin structure, which contributes to biological function.

AQ1

AQ2

AQ3

AQ4

AQ5

Results

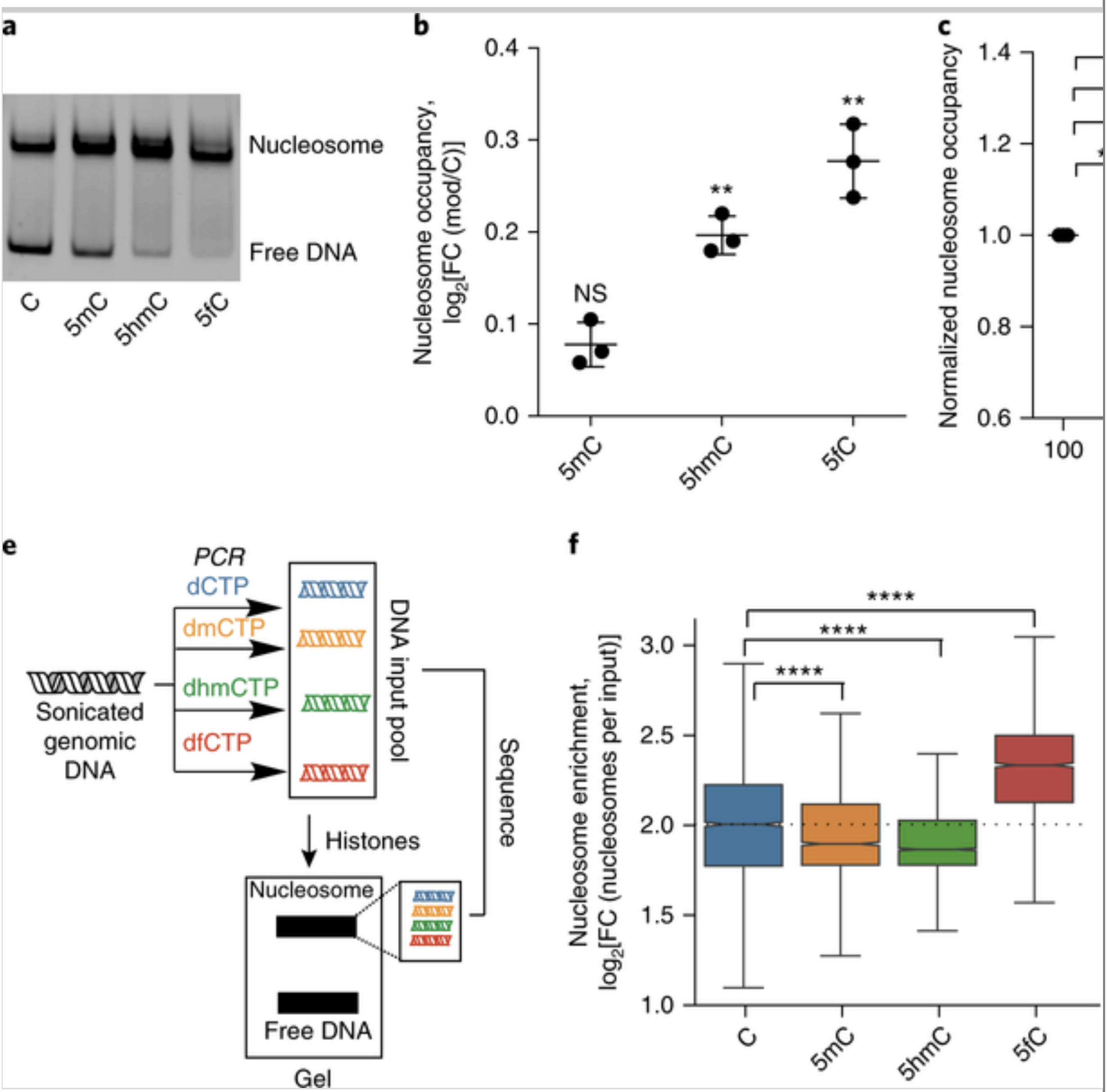
5fC increases nucleosome occupancy and stability

We assessed the differential effects of specific DNA modifications on nucleosome occupancy and stability using the Widom 601 DNA sequence. While previous studies have investigated the impact of cytosine-modified Widom 601 sequences on the conformation and stability of nucleosomes in vitro[4, 5, 6, 7, 16, 17, 18, 19, 20], here we sought to establish the relative strength of each cytosine modification to promote nucleosome occupancy. To address this, we used Cy-dye-labelled primers and PCR to generate the Widom 601 sequence bearing either cytosine, 5mC, 5hmC or 5fC in modified DNA strands that contained 67 base modifications, of which 10 were in a CpG context (Supplementary Table 1). After chaperone-mediated nucleosome assembly, we separated the free DNA from the nucleosome fraction by native gel electrophoresis (Fig.1a and Supplementary Fig.1). After quantification of both fractions we assessed the nucleosome occupancy (ratio of nucleosome to total DNA) for cytosine and modified cytosine DNA. We then calculated the $\log_2$fold changes in nucleosome occupancy and observed that 5mC, 5hmC and 5fC each significantly (two-tailed *t*-test, $P \leq 0.005$ for 5mC, $P \leq 0.0001$ for 5hmC and 5fC) caused an increase in nucleosome occupancy compared to unmodified Widom 601 DNA, with 5fC–DNA displaying the strongest effect (Fig.1b).

Fig. 1

5fC increases nucleosome occupancy and stability.

a, Double-stranded DNA comprising the Widom 601 sequence with cytosine, 5mC, 5hmC or 5fC was used to reconstitute nucleosomes. Cy3-labelled DNA was used for quantification. Non-denaturing polyacrylamide gel (4–16%) electrophoresis was run in 0.4× Tris/borate/EDTA (TBE) buffer to separate the nucleosome fraction from the free DNA. Experiments were repeated independently with similar results ($n = 3$). **b**, The $\log_2$ fold change (FC) of nucleosome occupancy (nucleosome/total DNA) of modified DNA compared to unmodified Widom sequence. Error bars represent s.d. of the mean from three independent experiments ($n = 3$). Two-tailed *t*-test was used to calculate *P* values (5mC, $P = 0.18$; 5hmC, $P = 3.3 \times 10^{-3}$; 5fC, $P = 6.8 \times 10^{-3}$). NS, not significant. **c**, Nucleosomes were reconstituted using different 5fC densities (100%, fully modified; 1%, approximately one 5fC per sequence). Normalized nucleosome fractions are plotted against 5fC density. Error bars represent s.d. of the mean from three independent experiments ($n = 3$). Two-tailed *t*-test was used to calculate *P* values (*$P = 0.0179$ and 0.0441, **$P = 0.001$). **d**, DNA–histone reconstitution was used for measuring the free energy of formation for 5fC nucleosomes. Differences in free energy changes ($\Delta\Delta G°$) were obtained by subtracting the free energy change for the unmodified Widom sequence from the free energy change for the 5fC-modified Widom sequence and are presented as the mean and s.d. of the mean from three independent experiments ($n = 3$). Two-tailed *t*-test was used to calculate *P* values (*$P = 0.0363$, 0.0113 and 0.0123, **$P = 0.0039$). **e**, A pool of DNA sequences containing either cytosine, 5mC, 5hmC or 5fC was formed by combining sublibraries, each generated by four parallel, indexed PCR reactions using genomic DNA. The pooled libraries were used for subsequent nucleosome reconstitution and sequencing. **f**, Nucleosome enrichment (nucleosome library/input DNA library) for cytosine (blue, $n = 2{,}977$), 5mC (orange, $n = 884$), 5hmC (green, $n = 10{,}537$) and 5fC (red, $n = 2{,}481$) DNA is plotted. Notched boxplot shows the first, second (median) and third quartiles, with whiskers extending to the minimum and maximum. *P* values were obtained using the two-sided Mann–Whitney test (****$P \leq 0.0001$: C versus 5mC, $P = 6 \times 10^{-9}$; C versus 5hmC, $P = 4.8 \times 10^{-16}$; C versus 5fC, $P = 7 \times 10^{-306}$). **g**, Representative genomic locus showing the enrichment of 5fC-associated nucleosomes (red) compared to cytosine (blue), 5mC (orange) and 5hmC (green) associated nucleosomes in reads per kilobase of transcript per million mapped reads (RPKM). Data are normalized to library input. Experiments were repeated twice independently with similar results ($n = 2$).



Next we assessed the lowest 5fC density that is sufficient to give a measurable effect on nucleosome occupancy. To this end, we generated Widom 601 sequences with 5fC densities ranging from fully modified (100%) down to approximately one 5fC unit per 147 bp DNA (1%) by changing the dfCTP/dCTP ratio during the PCR step (Supplementary Fig. 2 and Supplementary Table 2). Subsequent nucleosome assembly experiments revealed that only one unit of 5fC per nucleosome was required to increase nucleosome occupancy (Fig. 1c and Supplementary Fig. 3a). We also compared the free energy of DNA–histone interactions in nucleosomes as a function of 5fC density using the Widom 601 sequence (Fig. 1d and Supplementary Fig. 3b). For the measurement of relative affinities (free energy change) of histone–DNA interactions in nucleosomes we followed a previously published method from ref. [17]. Fluorescently labelled Widom 601 sequence was competed against a large excess of unlabelled competitor DNA (5S RNA sequence) for binding to the histone octamer, which was present in limited amounts. An initial NaCl concentration of 2 M was used to ensure equilibrium between the histone–DNA interactions, which was then gradually decreased to 125 mM by dialysis. The free energy was calculated from the equilibrium constant K_{eq} (fluorophore-labelled nucleosome fraction/fluorophore-labelled free DNA), which was obtained from quantification of the corresponding bands after native gel electrophoresis (see Methods for further details). Our data showed that decreasing 5fC density increased the free energy change for nucleosome formation when referenced to unmodified DNA, with 1% 5fC–DNA causing the most favourable shift ($\Delta\Delta G° = -0.286 \pm 0.02$ kcal mol^{-1}), consistent with our earlier band shift observations.

AQ6

Having demonstrated that 5fC, in particular, increased nucleosome occupancy for the Widom 601 sequence, we next investigated the preference for nucleosomes in a wider set of sequence contexts using a pool of genomic DNA sequences into which cytosine, 5mC, 5hmC or 5fC was introduced. To achieve this, we set up a competition assay for nucleosome formation with genomic DNA containing either 5fC, other C-modifications or no modifications. Starting from mouse genomic DNA, we generated four indexed DNA sublibraries comprising either no modifications or having each C position fully modified with either 5mC, 5hmC or 5fC. The sublibraries were generated by sonication of the genomic DNA to an average size of 150 bp, followed by adapter ligation and subsequent PCR amplification using either dCTP, d5mCTP, d5hmCTP or d5fCTP, besides dATP, dTTP and dGTP (Fig. 1e). These sublibraries of oligonucleotides were then pooled in equal amounts, to represent unmodified C and each C derivative in equal proportions, and the pool was subsequently used in a nucleosome reconstitution assay. Reconstituted nucleosomes were separated from free DNA on a non-denaturing gel and the nucleosome-bound DNA was then sequenced to identify the sequences and modifications that favour nucleosome formation. We accounted for coverage biases introduced during the initial PCR amplification by normalizing against the four input DNA libraries sequenced before nucleosome reconstitution. Modified bases in nucleosome DNA were identified by their index, and the enrichment of each modification was calculated as the $\log_2$ of the ratio between the sequencing coverage for nucleosome DNA and input DNA. Figure 1f shows the overall nucleosome enrichment for the different DNA libraries, revealing that of all the modifications evaluated, only 5fC-containing DNA was significantly (two-sided Mann–Whitney test, $P \leq 0.0001$) enriched in nucleosomes compared to unmodified DNA. Figure 1g presents an example of a genomic locus, where nucleosome signals are globally enriched in the 5fC-associated compared to cytosine, 5mC or 5hmC nucleosomes.

AQ7

5fC within its natural genomic sequence context enhances nucleosome occupancy in vitro

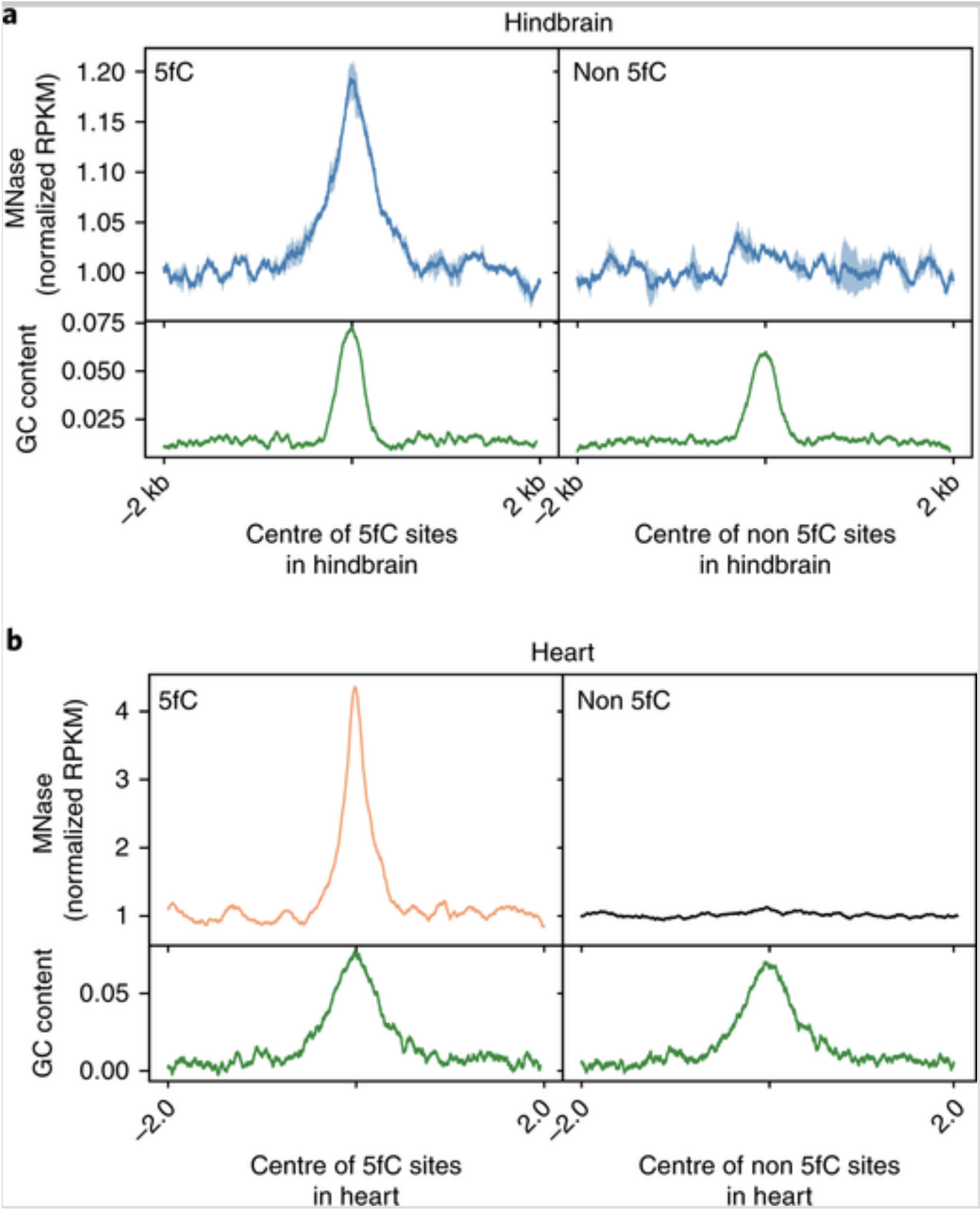
Given that genomic DNA where all Cs are replaced with 5fC causes an increase in nucleosome density, we next investigated if this observation still holds for the natural occurring levels of 5fC in natural sequence contexts. To accomplish this, we extracted genomic DNA from hindbrain and from heart tissues of E11.5 mouse embryos, then assembled the DNA into nucleosomes and subsequently sequenced the DNA after MNase treatment (Supplementary Table 3). This allowed us to identify DNA sequences bound by histone core proteins, which could be aligned to the 5fC maps in these tissues (data set taken from ref. [13]) to relate the presence of 5fC to the organization of nucleosomes. The data shown in Fig. 2a,b (upper panels) represent the averaged distribution of nucleosomes 2 kb upstream and downstream of 5fC sites previously identified in hindbrain (n = 7,114) and in heart (n = 1,080), respectively. These data show that natural 5fC sites have an increased nucleosome density, consistent with our in vitro observations (Fig. 1a–d,f). We considered first-order sequence context effects using genomic regions of 4 kb in length lacking 5fC, but with comparable average GC content, and then obtained the average nucleosome density for those regions (Fig. 2a,b, lower panels). Our analysis revealed that non-5fC sites with comparably high GC content did not show any increase in nucleosome density. Therefore, the link between nucleosomes and 5fC is independent of GC content. Collectively, these data demonstrate that 5fC in genomic DNA enhances nucleosome density.

Fig. 2

5fC within a genomic sequence context enhances nucleosome occupancy in vitro.

a,**b**, MNase reads (averaged normalized RPKM for hindbrain) of reconstituted nucleosomes using genomic DNA extracted from mouse embryonic hindbrain and heart (11.5 days) show enrichment around 5fC brain sites (blue) and 5fC heart sites (orange). Blue shading in **a** shows s.e.m. across biological replicates. MNase signals around non-5fC sites (black) with comparable GC content (green line) show depletion of MNase signal. As a control, we calculated the average genomic coverage on the same number of non-5fC sites (that is, sites that did not contain 5fC in any of the replicates). The non-5fC regions were randomly drawn from mappable genome sites such that their site-averaged CG enrichment profile across a 4,000 bp region matched that of the 5fC regions with a root-mean square error below 0.0025. Experiments were repeated twice, independently, with similar results ($n = 2$).

AQ8



5fC is a determinant of nucleosome organization in vivo that is linked to gene expression

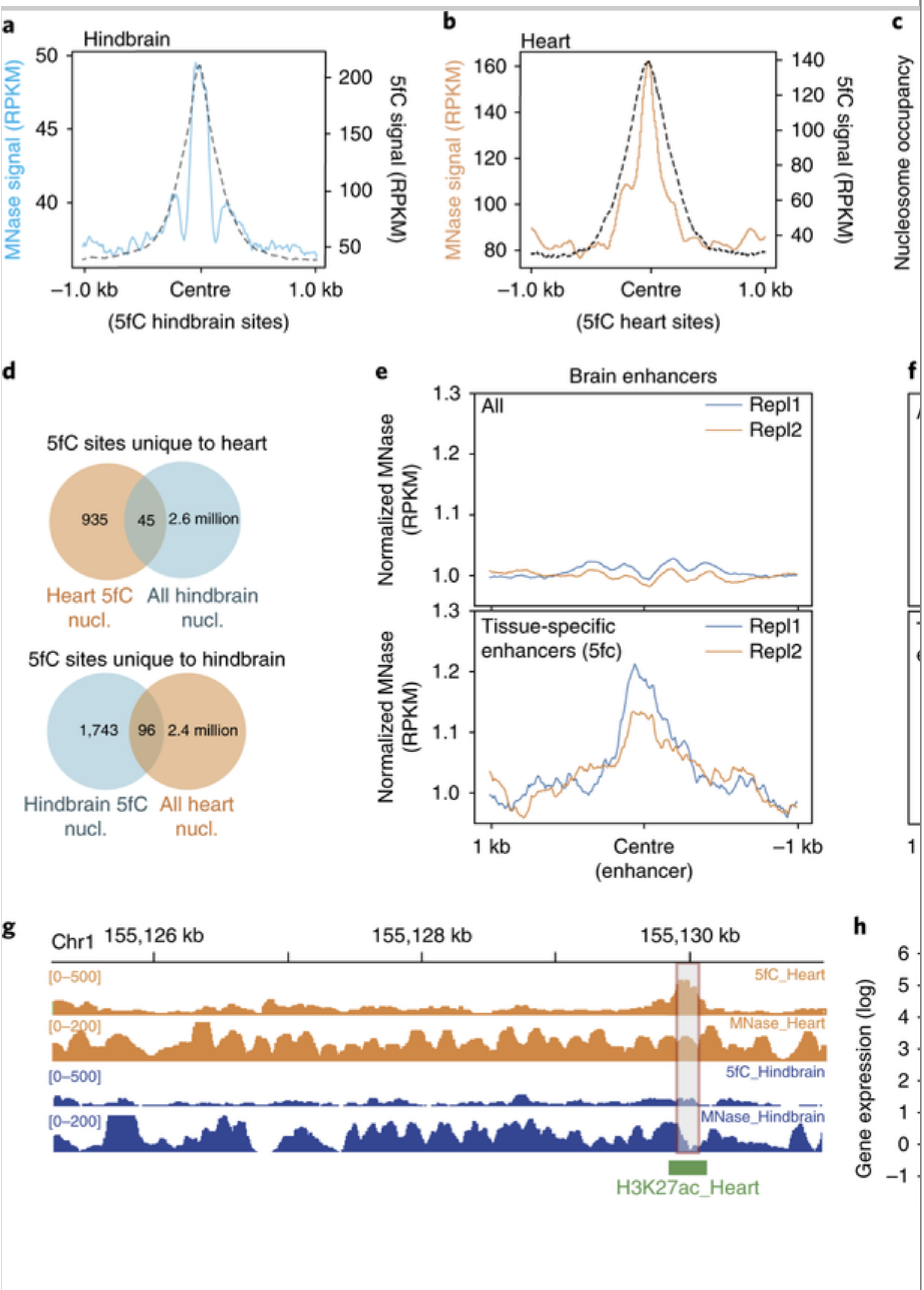
Because the locations of 5fC are tissue-specific[13] and 5fC levels are changing throughout development[12], there is the possibility that 5fC is involved in nucleosome organization that is tissue-specific. To address this possibility, we used MNase–seq to generate genome-wide nucleosome maps for the hindbrain and heart tissues of developing E11.5 mouse embryos (Supplementary Fig. 4 and Supplementary Table 3). We assessed the average nucleosome distribution 1 kb upstream and downstream of 5fC sites in hindbrain and heart. We observed increased nucleosome density centred at 5fC sites[13] in both hindbrain and heart tissues (Fig. 3a,b), consistent with our observations from in vitro nucleosome reconstitution assays. Notably, in hindbrain we observed well-positioned nucleosomes at the centre of 5fC sites, with flanking nucleosomes positioned adjacent to this site. Our data also revealed that, on average, the level of nucleosome occupancy at 5fC sites was significantly higher (two-sided Mann–Whitney *U*-test, $P \leq 0.0001$) than at all other nucleosomes located genome-wide at non-5fC sites, both in heart and hindbrain, demonstrating a stronger preference for nucleosomes to occupy 5fC–DNA sites in vivo (Fig. 3c).

Fig. 3

5fC is a determinant of nucleosome organization in vivo that is linked to gene expression.

a,**b**, Normalized MNase signal (RPKM) plotted around 5fC sites (dotted black line) in hindbrain (**a**, blue) and heart (**b**, orange) revealing increased nucleosome density at the centre of 5fC sites. **c**, Notched boxplot showing the nucleosome occupancy in hindbrain and heart tissues. Boxplot shows first, second (median) and third quartiles, with notches representing the confidence interval around the median. Average nucleosome occupancies within 5fC sites in heart (n = 11,226, orange) and hindbrain (n = 16,732, blue) are significantly higher ($P = 5.8 \times 10^{-17}$ for heart, $P < 3.3 \times 10^{-306}$ for hindbrain (below machine precision), two-sided Mann–Whitney *U*-test) than that of all detected nucleosomes in heart (n = 10,936,694) and hindbrain (n = 10,973,437). **d**, Venn diagram showing the overlap of nucleosomes at heart- and hindbrain-unique 5fC sites. **e**,**f**, Tissue-specific enhancers containing 5fC show differential nucleosome density compared to all enhancers. **g**, Representative genomic locus showing overlap between a heart-specific 5fC site and nucleosome as well as heart-specific H3K27ac in heart tissue that is absent in hindbrain. Experiments were repeated twice, independently, with similar results ($n = 2$). **h**, Comparison of gene expression (log RPKM) of a subset of enhancers marked by 5fC–nucleosomes (n = 517, light blue) and all enhancer sites (n = 13,286, dark blue) in hindbrain and 5fC–nucleosomes (n = 827, orange) and all enhancer sites (n = 13,286, red) in heart. Notched boxplot shows that the presence of 5fC at enhancer sites correlates with significantly higher gene expression ($P = 3.0 \times 10^{-8}$ for hindbrain, $P = 2.2 \times 10^{-4}$ for heart, two-sided Mann–Whitney *U*-test) of their associated genes compared to all expressed genes (predicted enhancer-gene list from ref. [23]). Boxplot shows first, second (median) and third quartiles, with notches representing the confidence interval around the median, the black diamond the mean, and whiskers the reach of data points beyond the first (Q1) and third (Q3) quartiles (for example, $Q1 + 1.5 \times (Q3 - Q1)$).

AQ9



When we compared in vitro MNase signals generated by nucleosome reconstitution with in vivo MNase–sequencing signals we observed higher

Pearson correlations of MNase signals at 5fC sites (Supplementary Fig. 5). At non-5fC-containing 5′-UTRs, exons and CGIs, for example, we observed no correlation between our in vitro and in vivo data sets; however, at 5fC-containing sites we observed increased correlation (r = 0.91) at 5fC-containing CGIs. This observation demonstrates an intrinsic 5fC–DNA preference of nucleosomes supporting a role for 5fC in determining the organization of nucleosomes.

AQ10

When we compared the nucleosome organization at 5fC sites between hindbrain and heart, we observed little overlap between nucleosomes at 5fC sites (Supplementary Fig. 6a). Notably, nucleosomes at 5fC sites unique to heart tissue had very little (<5%) overlap with genome-wide hindbrain nucleosomes, and vice versa, supporting a role for 5fC as a determinant of nucleosome organization that is tissue-dependent (Fig. 3d).

During development, the establishment of active, tissue-specific enhancers is important for the expression of genes that specify cell identity. It has been shown that tissue-specific enhancers exhibit relatively high nucleosome occupancy[21, 22] while allowing high accessibility, although it remains unclear how this differential nucleosome organization is established. Given our previous observation that 5fC was enriched at sites marked by H3K27ac and H3K4me1[13], both hallmarks of active enhancers, we next investigated whether 5fC contributes to the increased tissue-specific nucleosome density at enhancers. We therefore compared nucleosome densities across all enhancers to hindbrain- or heart-specific enhancers containing 5fC. We observed that enhancers were generally depleted of nucleosomes in hindbrain and heart (Fig. 3e,f, upper panels and Supplementary Fig. 6b). Tissue-specific enhancers containing 5fC, however, showed increased nucleosome densities, which is consistent with our hypothesis that 5fC promotes nucleosome formation (Fig. 3e,f, lower panels and Supplementary Fig. 6b). Figure 3g shows an example locus, where a heart-specific 5fC site overlaps with nucleosomes and a heart-specific H3K27ac site, whereas neither 5fC nor nucleosomes or H3K27ac are present in hindbrain tissue at this site. These data collectively support a role for 5fC in the organization of nucleosomes at regulatory regions that are important for defining hindbrain and heart tissues, respectively. This, in turn, suggests that 5fC-associated nucleosome organization could thus be linked to enhanced, tissue-specific gene expression. To evaluate this hypothesis, we generated RNA–seq data for hindbrain and heart and compared the expression of all genes that could be linked to predicted enhancers[23] against genes linked to predicted enhancers where we had detected 5fC-associated nucleosomes (Fig. 3h; see Supplementary Information for the definition of '5fC-associated nucleosomes'). We found that genes linked to 5fC-associated nucleosome enhancers were indeed significantly more highly expressed (two-sided Mann–Whitney U-test, $P \leq 0.0001$), supporting a role of 5fC in the organization of nucleosomes at regulatory sites that control gene expression.

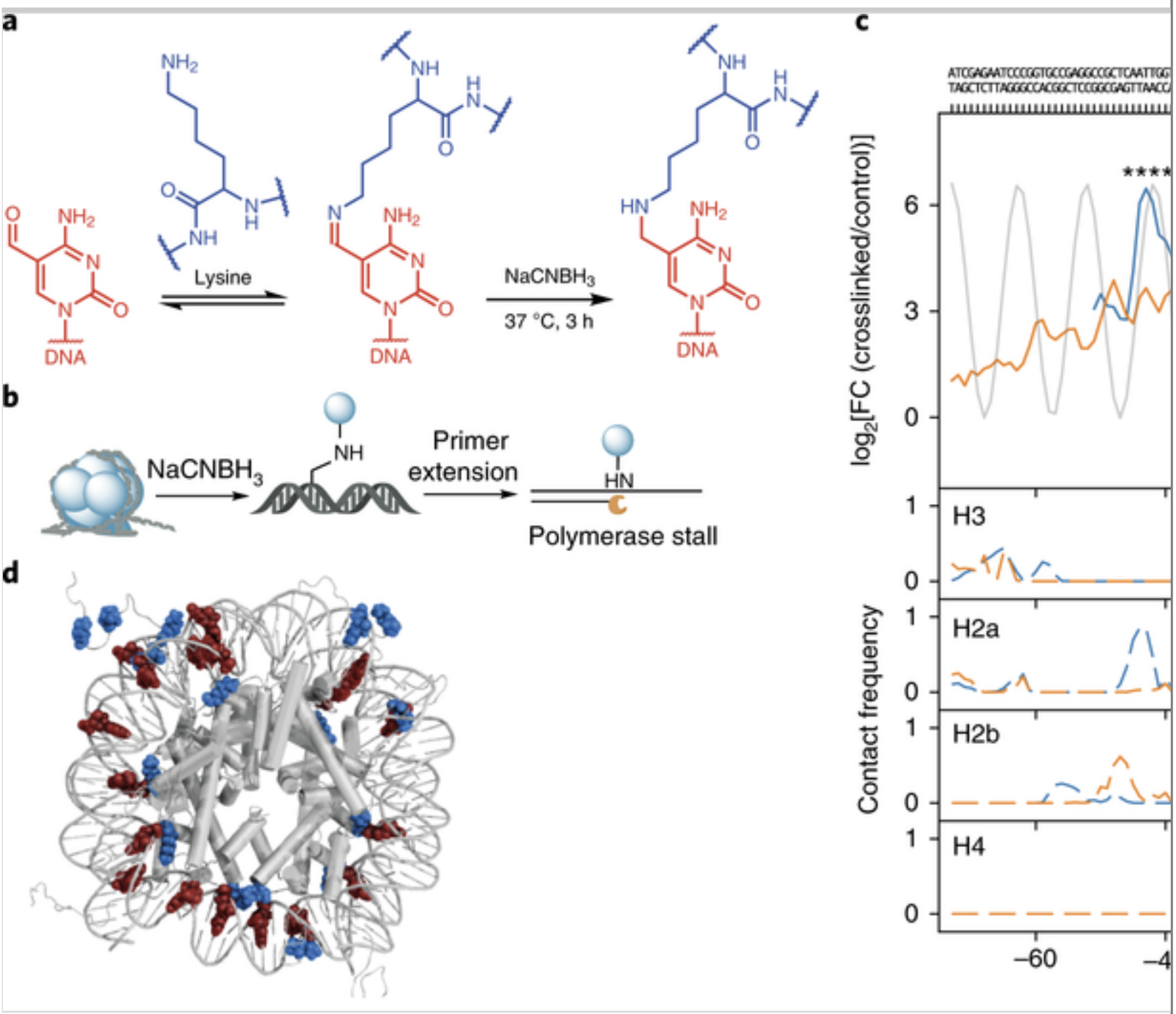
5fC–histone interaction model within a nucleosome

The aldehyde group in 5fC can react with primary amines of nuclear proteins (for example by the ε-amino group of lysine of histone proteins) to form a Schiff base (Fig. 4a)[14, 15]. To obtain insights into the explicit relationship of the Schiff base interaction between 5fC and lysine residues of the histone proteins within the nucleosome we created a structural interaction model. To do this we combined polymerase stalling experiments followed by sequencing to identify the 5fC-crosslinking sites, with molecular modelling to identify the critical, proximal lysine residues. Nucleosome core particles were assembled from the 5fC-containing Widom 601 sequence, then we used the reducing agent $NaCNBH_3$ to irreversibly trap any Schiff base, after its formation, by chemical reduction (Supplementary Fig. 7). The crosslinking event was confirmed by denaturing gel electrophoresis, which showed the appearance of a new higher-molecular-weight band that was not observed in the control DNA, which lacked 5fC (Supplementary Fig. 8). We cannot rule out the possibility that irreversible trapping by chemical reduction may shift the equilibrium to the formation of a Schiff base, thereby potentially leading to overestimation of the covalent histone–DNA interaction. For the polymerase stalling experiment we carried out a single primer extension reaction on the 5fC–nucleosome DNA, after $NaCNBH_3$ reduction, followed by sequencing to identify stalling sites (Fig. 4b and Supplementary Fig. 9a). Significant stalling sites (false discovery rate (FDR) < 0.0001, Benjamini–Hochberg correction on the exact P value from the negative binomial distribution) were assessed by comparing the data set against the negative control, where only free 5fC-containing DNA (in the absence of any histone proteins) was used for the primer extension step to account for natural polymerase stalling events. When we computed the moving average of $\log_2$ fold change stalling sites for the forward and reverse Widom 601 template, we observed a stalling pattern with distinct ~10 bp periodicity that was in phase with the major groove of DNA facing the histone core (Fig. 4c). Our analysis revealed that all stalling maxima had 5fC either directly at or immediately before the stalling site, indicating pronounced polymerase stalling at 5fC-crosslinked sites. In particular, we observed significant stalling sites where 5fC was in a CpG context (highlighted in grey within the sequence in Fig. 4c), demonstrating that 5fC specifically at CpG sites, rather than in other contexts, engaged in covalent Schiff base interactions between histones and DNA.

Fig. 4

5fC–histone interaction model within a nucleosome.

a, Scheme showing the chemistry of the irreversible Schiff base trap. **b**, DNA polymerase stalling sites after $NaCNBH_3$ treatment were identified by sequencing to reveal the sites of 5fC–histone covalent bonds around the nucleosome core. **c**, Top, $\log_2$ fold change between the number of reads of crosslinked and control sample for the forward (blue) and reverse strand (orange). The grey sinusoidal line indicates the orientation of the major groove with respect to the histone core, ranging from 1 (facing the histone core) to −1 (facing away from the histone core). Significant (FDR values of 9.6×10^{-5}, 9.6×10^{-5}, 1.1×10^{-4} and 9.6×10^{-5} from left to right, Benjamini–Hochberg correction on the exact P value from the negative binomial distribution) stalling sites (±3 bases) around CpG dinucleotides within the DNA sequence are highlighted in the DNA sequence in grey. Bottom, The closest lysines (<5 Å) in H2A, H2B, H3 and H4 facing the major grooves of the DNA pointing towards the histone core were identified. Based on the overlap of the computed data with stalling sites, potential 5fC sites involved in Schiff base formation were identified. Experiments were repeated twice independently with similar results ($n = 2$). **d**, Structure highlighting 5fC sites (red balls) within the DNA and lysine residues (blue balls) that potentially interact through the formation of Schiff base conjugates.



Next, we combined the data from polymerase stalling with molecular modelling to identify the key lysine residues proximal to these stalling sites. We extracted the proximal lysine-nucleotide distances (within 5 Å) for all the bases along the Widom 601 nucleosome from molecular dynamics simulations (Supplementary Fig. 9b), and quantified their contact frequencies (the fraction of time a pair of atoms is less than 5 Å apart; see Supplementary Information for details). The contact frequencies between the bases and lysine side chains of histone proteins H2a/b, H3 and H4 are depicted in Fig. 4c (bottom panels), showing that lysine residues from all four histone proteins are in close proximity to significant stalling sites. Notably, histone H3 displayed a high contact frequency around the 5fC sites within a CpG context. Table 1 and Fig. 4d summarize our findings from polymerase stalling and molecular modelling (predicted lysine residues are visualized in blue in the crystal structure used for the modelling with proximal 5fC sites coloured in red).

Table 1

Details of the crosslinked 5fC position, the proximal lysine histone subunit and whether lysine is part of the core histone subunit or histone tail

5fC residue	Lysine residue	Histone subunit
26 (fw)	2	H2B_1 Tail
36 (fw)	8	H2A_1 Tail
47 (fw)	12	H2B_2 Tail
56 (fw)	11	H4_1
74 (fw)	115	H3_1
84 (fw)	31	H4_2
96 (fw)	16	H4_2
111 (rev)	122	H2B_2
22 (rev)	24	H2B_1 Tail
33 (rev)	36	H2A_1 Tail
50 (rev)	12	H2B_2 Tail
63 (rev)	23	H3_1
73 (rev)	115	H3_1
85 (rev)	31	H4_2
94(rev)	16	H4_2

AQ11

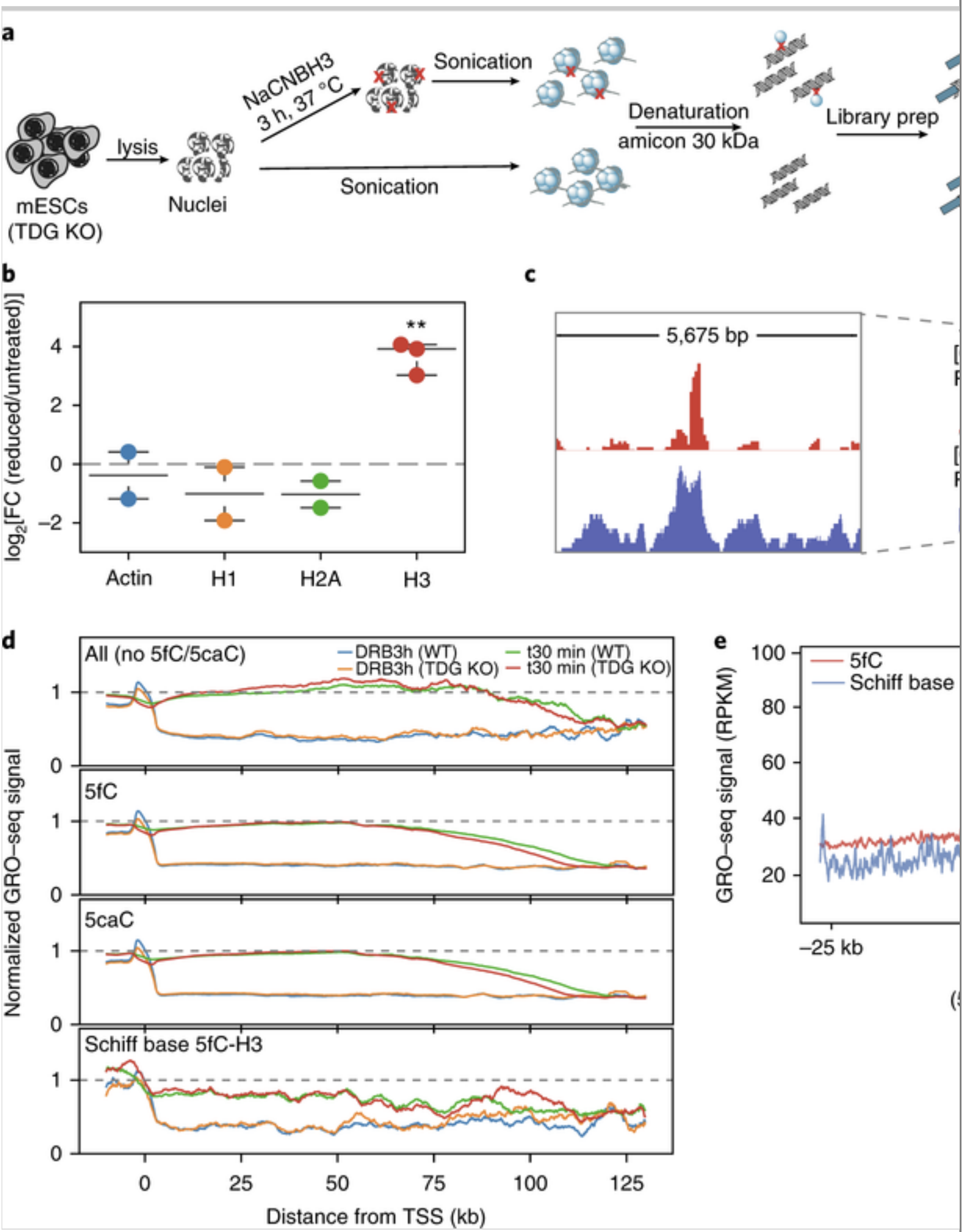
5fC can form a Schiff base with histones in chromatin

We next explored the existence of Schiff base interactions between the 5fC of genomic DNA and the lysines of histone proteins in the chromatin of mESCs, to gain insights into the biological relevance of this interaction (Fig. 5a). We used $NaCNBH_3$ to chemically trap any Schiff base formed in the nuclei from TDG knockout (KO) mESCs that contain relatively high levels of 5fC[24]. We considered that $NaCNBH_3$ can also trap adducts formed by lysine residues reacting with naturally occurring DNA abasic sites. However, most polymerases cannot amplify DNA containing abasic sites[25, 26]. Our quantitative PCR (qPCR) analysis on synthetic DNA containing either 5fC or an abasic site with or without a crosslinked lysine residue confirmed that the polymerase was not able to efficiently amplify DNA with abasic sites (with or without crosslinked lysine), whereas 5fC-containing DNA (with or without crosslinked lysine) was efficiently amplified (Supplementary Fig. 10).

Fig. 5

5fC can form a Schiff base with histones in a chromatin context that impacts transcription elongation.

a, Workflow for the detection of in vivo Schiff base sites in TDG KO mESCs. Key steps involve the reduction of the imine bond using $NaCNBH_3$ followed by denaturation of proteins to disrupt any non-crosslinked DNA–protein interaction, and subsequent histone ChIP, with no reduction for the control. **b**, $\log_2$fold change of reduced/untreated samples after ChIP–qPCR, showing significant enrichment (two-tailed *t*-test, $P = 7.7 \times 10^{-3}$) of H3 immunoprecipitated chromatin after reduction. The scatter dot plot shows the values for individual replicates for actin (blue), H1 (orange), H2A (green) and H3 (red) with lines indicating the mean with s.d. Experiments were repeated twice for actin, H1 and H2A and three times for H3, independently, with similar results. **c**, Representative genomic locus showing the overlap between 5fC (red) and crosslinked H3 sites (blue) in TDG KO mESCs. Experiments were repeated twice with similar results. **d**, Metagene analysis of normalized GRO–seq signal at 'all' (without 5fC/5caC) genes and genes containing 5fC, 5caC or Schiff base 5fC/H3 sites. **e**, GRO–seq signals (RPKM) centred around 25 kb up- and downstream of Schiff base 5fC–H3 sites (blue) and 5fC only (red) sites.



AQ12

After $NaCNBH_3$ reduction, chromatin was sonicated to an average size of 150 bp. Non-covalent DNA–protein interactions were disrupted by guanidine HCl treatment followed by extensive washing to generate histone–DNA conjugates in the reduced but not untreated sample, owing to Schiff base formation. After end repair, A-tailing and adapter ligation, we isolated covalent histone–DNA conjugates by histone chromatin immunoprecipitation (ChIP) using four different antibodies to H1, H2A, H3 and actin, respectively. Subsequent proteinase K treatment, to digest the proteins, followed by qPCR using universal Illumina primers, allowed us to quantify the enrichment of DNA libraries of the reduced ($NaCNBH_3$-treated) sample obtained by ChIP compared to the unreduced sample. We observed significant enrichment (two-tailed *t*-test, $P = 7.7 \times 10^{-3}$) when the antibody for H3 was used, but not for actin, H1 or H2A, suggesting covalent Schiff base formation primarily between 5fC and H3 (Fig. 5b). We then sequenced two biological replicates of the H3 ChIP libraries to identify the sites of Schiff base formation (Supplementary Table 3). We found a total of 1,461 peaks (union across two replicates), which were then cross-correlated with existing 5fC maps[27] to identify the sites of covalently 5fC-bound histones in the genome. We found 364 sites that overlapped with 5fC sites, half of which were found within genes (164 genes). As determined using simulated random distribution, these sites showed significant ($P < 0.0001$) overlap with 5fC sites. Figure 5c shows a representative locus where a covalently linked 5fC–H3 site was identified within the gene of Ptpn14.

Nascent RNA analysis can provide insights into the dynamics of transcription by measuring how far transcription by RNA Pol II proceeds from the transcription start site during a given time window, capturing the so-called transcription wave. As a step towards understanding whether naturally occurring covalent 5fC–H3 Schiff base interactions have a direct effect on the mechanism of transcription, we analysed the transcription wave of newly synthesized (nascent) RNA transcripts using a pre-existing Global Run On–Sequencing (GRO–seq) data set[28]. We studied nascent RNA elongation on wild-type (WT) mESCs and TDG KO mESCs. Our analysis revealed that genes with 5fC (but no Schiff base) or 5caC showed retarded RNA Pol II elongation in TDG KO relative to WT as measured by the differences in transcription elongation (Fig. 5d and Supplementary Fig. 11). This observation supports earlier reports that 5fC and 5caC reduce the rate of RNA Pol II[28, 29]. For genes in which we have detected the Schiff base 5fC–H3 conjugate we observed a decay in transcription elongation, as measured by the loss of GRO–seq signal just after the transcription start sites, concomitant with a new wave of nascent transcription activity after 75 kb. This effect was particularly pronounced in TDG KO. Furthermore, an analysis of the density of nascent RNA sequencing reads around Schiff base 5fC–H3 sites revealed a peak in the GRO–seq signal ~2 kb downstream of the Schiff base sites that was not observed at 5fC (without Schiff base) sites, demonstrating the presence of transcriptionally active Pol II immediately downstream of Schiff base sites (Fig. 5e).

Discussion

We used synthetic and genomic DNA to investigate if nucleosomes exhibited a preference for unmodified, 5mC, 5hmC or 5fC-containing DNA. Our data revealed that, while 5mC, 5hmC and 5fC all showed increased nucleosome occupancy compared to unmodified cytosine when the synthetic Widom DNA sequence was used, 5fC in particular also caused strong promotion of nucleosome occupancy within the genomic DNA sequence context. Our results suggest that the preference of nucleosomes for 5fC-containing DNA is largely independent of the sequence context. In contrast, we observed that the increased nucleosome occupancy observed with methylated or hydroxymethylated Widom sequence was rather sequence context specific, because these modifications within the genomic DNA sequence context were generally linked to decreased nucleosome occupancy. This was further supported by in vivo data showing that naturally occurring 5fC, within its natural genomic context, occurred at loci that have increased nucleosome occupancy and that contribute to tissue-specific nucleosome organization and gene expression. Previous studies have shown that regulatory regions including active enhancers are generally nucleosome depleted to ensure accessibility to regulatory proteins[1, 30]. In contrast, some active tissue-specific enhancers exhibit relatively high nucleosome occupancy and accessibility, thereby supporting a model where tissue-specific gene regulation is facilitated by nucleosome-mediated pioneer transcription factor activity[22, 31]. Our data suggest that 5fC contributes to the differential nucleosome organization at tissue-specific enhancers that are also linked to increased expression of the associated genes. Although we cannot exclude that the nucleosome organization at these regulatory sites is influenced by 5mC or 5hmC, earlier genome-wide studies in ESCs and mammalian tissues have identified patterns of 5fC that were distinct from

5hmC and 5mC[32, 33, 34]. These studies have shown that 5hmC is enriched at promoters, gene bodies and poised enhancers, which is distinct from 5fC sites, which are enriched at active enhancers in embryonic mouse tissues.

Our 5fC–histone interaction model provides a structural explanation for how 5fC-containing DNA can stabilize and position nucleosomes. A very recent report provided evidence that Schiff base interactions can form between 5fC and lysines in cells[14], but that study did not identify the proteins from which the key lysine(s) originated. By chemically trapping 5fC–lysine conjugates before immunoprecipitation using antibodies against histone proteins, we now provide evidence that 5fC–histone(H3) conjugates can occur in mammalian cells. Notably, our study revealed that in mESCs, 5fC–H3 covalent interactions affect Pol II transcription elongation rates and also mark sites of active Pol II. This situation bears similarity to the proximal pausing of active Pol II just downstream of transcription start sites, which is also characterized by a burst of nascent transcript[35]. Polymerase stalling events caused by covalent 5fC–histone interactions may provide an opportunity for the recruitment of proteins involved in transcription regulation and could represent a key regulatory step in the control of gene expression by 5fC.

Collectively, our data support a model whereby 5fC contributes to the organization of cell and tissue-specific nucleosomes, providing a molecular mechanism to help explain how 5fC regulates gene expression during development and how it may be involved in the reinforcement of cell identity.

Methods

Chaperone-assisted nucleosome assembly

Master mix was prepared following the manufacturer’s instructions (Chromatin Assembly Kit, Active Motif Belgium) with some modifications. Briefly, for 100 ng DNA assembly, 1.5 µl high salt buffer was incubated with 0.21 µl h-NAP-1 and 0.27 µl HeLa core histones. The mixture was incubated on ice for 30 min before the addition of 3.65 µl low salt buffer, 0.38 µl ACF complex and 1.5 µl freshly prepared complete 10× ATP regeneration system. Complete 10× ATP regeneration system was prepared by mixing 0.1 μl creatine kinase with 1.65 μl 10× ATP regeneration system. The mixture was gently agitated after the addition of each component. DNA (100 ng) was diluted with ultrapure water to 7.5 μl and mixed with 7.5 μl master mix before incubation at 27 °C (block) with 50 °C (lid) overnight. Finally, gel electrophoresis was performed for quantification of the nucleosome fraction and free DNA.

AQ13

In vitro reconstitution of chromatin

Genomic DNA was extracted from mouse embryonic tissues at 11.5 days using the DNeasy Blood & Tissue kit from Qiagen.

Hindbrain/heart genomic DNA reconstitution

For 500 ng genomic DNA nucleosome reconstitution experiments, the master mix for nucleosome reconstitution was prepared as described above with small changes. NAP1 (1.4 μl), HeLa histone (1.8 μl) and high salt buffer (10 μl) were mixed and incubated on ice for 15 min. Low salt buffer (64.3 μl) was added together with the ACF complex (2.5 μl), 10× ATP regeneration system (10 μl), 500 ng DNA and water, to a final volume of 100 μl. Two biological replicates were generated for each condition—'treated' and 'untreated'—for subsequent MNase sequencing.

Cytosine or modified cytosine genomic DNA reconstitution

Genomic DNA (from mouse embryonic hindbrain tissues) was sonicated to 150 bp average size, end repaired, A-tailed, and ligated to indexed Illumina adapters using the standard Illumina library preparation method (NEBNext Ultra II DNA library Prep). For the indexed library, cytosine or fully modified 5mC, 5hmC or 5fC–DNA was subsequently generated by PCR (25 cycles) using Taq polymerase and purified by the GeneJET PCR Purification Kit and eluted with ultrapure water. 100 ng of each indexed library was pooled (400 ng total DNA) and used for nucleosome reconstitution following the chaperone-assisted assembly protocol. After incubation with histone octamer (octamer/DNA, 1:0.75) for 16 h at 27 °C, the nucleosome fraction was separated from free DNA on a 6% DNA retardation gel. The nucleosome band was cut out and soaked in 200 μl of 300 mM NaOAc and 1 mM EDTA (pH 8) for 48 h. After passing through a Spin-X centrifuge filter (pore size 0.22 μm, Sigma Aldrich), the supernatant was washed twice through an Amicon Ultra spin column (10 kDa cutoff, Merck). Input DNA libraries (before nucleosome reconstitution) and nucleosome fraction libraries were PCR amplified for six cycles with unmodified dNTPs, and sequenced on an Illumina NextSeq 500. Two replicates were generated for each modification.

In vivo Schiff base detection by qPCR followed by sequencing

To prepare cell nuclei, 2 ml of 2× lysis buffer (20 mM Tris pH 7.4, 10 mM $MgCl_2$, 2% Triton X-100 and 0.65 M sucrose) were added to a cell pellet (10 million mESCs) in 2 ml PBS. After incubation for 5 min on ice, nuclei were split into 2 × 2 ml suspensions (one control and one reduced sample) and repelleted for 15 min at 4 °C (4,000 r.p.m.). For the chemical reduction, 2 ml of 80 mM $NaCNBH_3$ of PBS was added to the nuclei and incubated for 3 h at 37 °C. Subsequently, untreated and reduced nuclei were sonicated for 8 min on a Covaris M220 using the a truChIP Chromatin Shearing Kit (Covaris) following the manufacturer’s guidelines (at this point the size distribution of sonicated chromatin was assessed by tape station after Genejet purification of a small aliquot). Protein denaturation was achieved using binding buffer containing a chaotropic agent (from the Genejet PCR purification kit, Thermofisher) followed by 4 × 450 µl water washes in the Amicon Ultra spin column (30 kDa cutoff). After end repair, A-tailing and adapter ligation (no purification step), DNA libraries were used for ChIP. For this, 100 µl Dynabeads, M-280 sheep anti-rabbit IgG, were washed and pre-incubated with 10 µg antibodies (Upstate 06597, ab18255, ab1791 and ab8227) in PBS/BSA for 3 h at 4 °C. Beads were subsequently washed with PBS/BSA, and DNA libraries together with 1 µg salmon sperm DNA were added to the beads and incubated overnight at 4 °C. Beads were then washed (5 min rotation at room temperature) six times with 500 µl LiCl buffer (100 mM Tris pH 7.4, 500 mM LiCl, 1% NP40, 1% sodium deoxycholate), with three tube changes. For elution, beads were incubated (15 min rotation at room temperature) with 2 × 100 µl elution buffer (0.1 $NaHCO_3$, 1% SDS). To the combined 200 µl supernatant were then added 8 µl of 5 M NaCl, 4 µl of Tris pH 7.4, 2 µl of 0.5 M EDTA and 1 µl proteinase K (10 mg ml^{-1}) and incubated at 45 °C for 1 h. After purification using the Genejet PCR purification kit, enrichments of the reduced over untreated samples were assessed using Illumina PCR primers and the Kapa quantification kit (Kapa Biosystems). DNA libraries were amplified (16 cycles) and subsequently sequenced.

Bioinformatic analysis

The methods used for the bioinformatics analysis are described in the Supplementary Information.

Code availability

The scripts to process the raw data associated with the manuscript, as well as custom computer code required to reproduce our results, are available from https://github.com/slab-bioinformatics/5fC-nucleosome.

Ethics statement

All experimental procedures were approved by the Animal Welfare, Experimentation and Ethics Committee at the Babraham Institute and were performed under licences from the Home Office (UK) in accordance with the Animals (Scientific Procedures) Act 1986.

Electronic supplementary material

Supplementary information is available for this paper at https://doi.org/10.1038/s41557-018-0149-x.

Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Author contributions

E.A.R., S.B., W.R. and M.A.D. designed the study. E.A.R., R.H., Z.L., W.D., M.I., J.S. and Z.L. performed the experiments. G.P., S.M.C. and D.B. performed the computational analysis. All authors analysed and interpreted the data. E.A.R., G.P. and S.B. wrote the manuscript, with contributions from all authors.

AQ14

Data availability

Sequencing data are available in the ArrayExpress database (www.ebi.ac.uk/arrayexpress) under accession no. E-MTAB-6271. 5fC regions in tissues are available in the GEO database under accession no.

GSE77447. 5fC and 5caC regions in mESCs are available under accession no. GSE42250. Histone marks were obtained from ENCODE (https://www.encodeproject.org/) with the following accession numbers: H3K27ac (ENCFF203QTV) and H3K4me1 (ENCFF542GAS) for hindbrain; H3K27ac (ENCFF954URD) and H3K4me1 (ENCFF737FNO) for heart. Data sets corresponding to the GRO–seq experiments were obtained from the GEO database under accession no. GSE64748.

Competing interests S.B. is a founder, advisor and shareholder of Cambridge Epigenetix Ltd.

Supplementary information

Supplementary Information

Supplementary Methods, Supplementary Figures, and Supplementary Tables

Reporting Summary

References

1. Struhl, K. & Segal, E. Determinants of nucleosome positioning. *Nat. Struct. Mol. Biol.* **20**, 267–273 (2013).

2. Kaplan, N. et al. The DNA-encoded nucleosome organization of a eukaryotic genome. *Nature* **458**, 362–366 (2009).

3. Segal, E. et al. A genomic code for nucleosome positioning. *Nature* **442**, 772–778 (2006).

4. Mendonca, A., Chang, E. H., Liu, W. & Yuan, C. Hydroxymethylation of DNA influences nucleosomal conformation and stability in vitro. *Biochim. Biophys. Acta* **1839**, 1323–1329 (2014).

5. Jimenez-Useche, I. et al. DNA methylation regulated nucleosome

dynamics. *Sci. Rep.* **3**, 2121 (2013).
AQ15

6. Choy, J. S. et al. DNA methylation increases nucleosome compaction and rigidity. *J. Am. Chem. Soc.* **132**, 1782–1783 (2010).

7. Lee, J. Y. & Lee, T.-H. Effects of DNA methylation on the structure of nucleosomes. *J. Am. Chem. Soc.* **134**, 173–175 (2012).

8. Teif, V. B. et al. Nucleosome repositioning links DNA (de)methylation and differential CTCF binding during stem cell development. *Genome Res.* **24**, 1285–1295 (2014).

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

10. 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).

11. Wu, X. & Zhang, Y. TET-mediated active DNA demethylation: mechanism, function and beyond.*Nat. Rev. Genet.* **18**, 517–534 (2017).
AQ16

12. Bachman, M. et al. 5-Formylcytosine can be a stable DNA modification in mammals. *Nat. Chem. Biol.* **11**, 555–557 (2015).

13. Iurlaro, M. et al. In vivo genome-wide profiling reveals a tissue-specific role for 5-formylcytosine. *Genome Biol.* **17**, 141 (2016).

14. Ji, S., Shao, H., Han, Q., Seiler, C. L. & Tretyakova, N. Y. Reversible DNA–protein cross-linking at epigenetic DNA marks. *Angew. Chem. Int. Ed.* **56**, 14130–14134 (2017).

15. Li, F. et al. 5-Formylcytosine yields DNA–protein cross-links in nucleosome core particles. *J. Am. Chem. Soc.* **139**, 10617–10620 (2017).

16. Ngo, T. T. M. et al. Effects of cytosine modifications on DNA flexibility and nucleosome mechanical stability. *Nat. Commun.* **7**, 10813 (2016).

17. Thåström, A., Lowary, P. & Widom, J. Measurement of histone–DNA interaction free energy in nucleosomes. *Methods* **33**, 33–44 (2004).

18. Lowary, P. & Widom, J. New DNA sequence rules for high affinity binding to histone octamer and sequence-directed nucleosome positioning. *J. Mol. Biol.* **276**, 19–42 (1998).

19. Dong, F., Hansen, J. C. & van Holde, K. E. DNA and protein determinants of nucleosome positioning on sea urchin 5S rRNA gene sequences in vitro. *Proc. Natl Acad. Sci. USA* **87**, 5724–5728 (1990).

20. Frouws, T. D., Duda, S. C. & Richmond, T. J. X-ray structure of the MMTV-A nucleosome core. *Proc. Natl Acad. Sci. USA* **113**, 1214–1219 (2016).

21. Tillo, D. et al. High nucleosome occupancy is encoded at human regulatory sequences. *PLoS ONE* **5**, e9129 (2010).

22. Iwafuchi-Doi, M. et al. The pioneer transcription factor FoxA maintains an accessible nucleosome configuration at enhancers for tissue-specific gene activation. *Mol. Cell* **62**, 79–91 (2016).

23. Shen, Y. et al. A map of the *cis*-regulatory sequences in the mouse genome. *Nature* **488**, 116–120 (2012).

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

25. Haracska, L., Washington, M. T., Prakash, S. & Prakash, L. Inefficient bypass of an abasic site by DNA polymerase η. *J. Biol. Chem.* **276**, 6861–6866 (2001).

26. Hogg, M., Wallace, S. S. & Doublié, S. Crystallographic snapshots of a replicative DNA polymerase encountering an abasic site. *EMBO J.*

23, 1483–1493 (2004).

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

28. Wang, L. et al. Molecular basis for 5-carboxycytosine recognition by RNA polymerase II elongation complex. *Nature* **523**, 621–625 (2015).

29. Kellinger, M. W. et al. 5-Formylcytosine and 5-carboxylcytosine reduce the rate and substrate specificity of RNA polymerase II transcription. *Nat. Struct. Mol. Biol.* **19**, 831–833 (2012).

30. West, J. A. et al. Nucleosomal occupancy changes locally over key regulatory regions during cell differentiation and reprogramming. *Nat. Commun.* **5**, 4719 (2014).

31. Mieczkowski, J. et al. MNase titration reveals differences between nucleosome occupancy and chromatin accessibility. *Nat. Commun.* **7**, 11485 (2016).

32. Sérandour, A. A. et al. Dynamic hydroxymethylation of deoxyribonucleic acid marks differentiation-associated enhancers. *Nucleic Acids Res.* **40**, 8255–8265 (2012).

33. Wen, L. et al. Whole-genome analysis of 5-hydroxymethylcytosine and 5-methylcytosine at base resolution in the human brain. *Genome Biol.* **15**, R49 (2014).

34. Song, C.-X. et al. Selective chemical labeling reveals the genome-wide distribution of 5-hydroxymethylcytosine. *Nat. Biotechnol.* **29**, 68–72 (2011).

35. Core, L. J., Waterfall, J. J. & Lis, J. T. Nascent RNA sequencing reveals widespread pausing and divergent initiation at human promoters. *Science* **322**, 1845–1848 (2008).