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

Eun-Ang Raiber[1,9], Guillem Portella[1,9], Sergio Martínez Cuesta[1,2], Robyn Hardisty[1], Pierre Murat[1], Zhe Li[1], Mario Iurlaro[3,4], Wendy Dean[3], Julia Spindel[3], Dario Beraldi[5], Zheng Liu[1], Mark A. Dawson[6], Wolf Reik[3,7], Shankar Balasubramanian[1,2,8,*]

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

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

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

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

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

[9]These authors have contributed equally to this work

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

Abstract

Nucleosomes are the basic unit of chromatin that help the packaging of genetic material whilst controlling access to the genetic information. The underlying DNA sequence, together with transcription-associated proteins and chromatin remodeling complexes, are important factors that influence the organization of nucleosomes. Herein, 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 is 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.

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[1–3], *in vitro* studies have shown that DNA base modifications have the potential to alter the conformation and stability of nucleosomes[4–7].

Other studies in mESC, have linked the presence of 5-hydroxymethylcytosine (5hmC) to the depletion of nucleosome occupancy, a correlation that was not observed for 5fC.[8] 5-Formylcytosine is a natural, modified base that is generated by the oxidation of 5-methylcytosine (5mC) by TET enzymes.[9] The 5fC base can be removed *via* base excision repair catalyzed 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 5-hydroxymethylcytosine (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. Herein, 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 mouse embryonic stem cells. Our data support the existence of a molecular linkage between DNA base modification and chromatin structure, which contributes to biological function.

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. Whilst previous studies have investigated the impact of cytosine modified Widom 601 sequences on the conformation and stability of nucleosomes *in vitro*[4–7,16–20] , here we sought to establish the relative strength of each cytosine modification to promote nucleosome occupancy. To address this, we used Cy-dye labeled 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 log2-fold changes in nucleosome occupancy and we observed that 5mC, 5hmC and 5fC each significantly (two-tailed t-test, p-value≤ 0.005 for 5mC, p-value ≤ 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).

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 ~1 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 1 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 the 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 by Thåström et al.[17] Therefore, fluorescently labeled Widom 601 sequence was competed against a large excess of unlabeled competitor DNA (5S RNA sequence) for the binding to 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 Keq (fluorophore-labeled nucleosome fraction/ fluorophore-labeled free DNA), which was obtained from quantification of the corresponding bands after native gel electrophoresis (see Method 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 favorable shift ($\Delta\Delta G° = -0.286 \pm 0.02$ kcal mol^{-1}), consistent with our earlier band shift observations.

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 were 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 sub-libraries comprising either no modifications or having each C-position fully modified with either 5mC, 5hmC or 5fC. The sub-libraries 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 sub-libraries 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 log2 of the ratio between the sequencing coverage for nucleosome DNA and input DNA. Fig. 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-value ≤ 0.0001) enriched in nucleosomes compared to unmodified DNA. Fig. 1g represents an example of a genomic locus, where nucleosome signals are globally enriched in the 5fC-associated compared to cytosine, 5mC or 5hmC nucleosomes.

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 (dataset taken from Iurlaro et al.[13]) to relate the presence of 5fC to the organization of nucleosomes. The data shown in Fig. 2a and 2b (upper panels) represent the averaged distribution of nucleosomes 2kb upstream and downstream of 5fC sites previously identified in hindbrain (n = 7114) and in heart (n = 1080), respectively. These data show that natural 5fC sites have an increased nucleosome density, consistent with our *in*

vitro observations (Fig 1a-d, Fig 1f). 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 and 2b, 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.

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

Since the locations of 5fC are tissue-specific[13] and 5fC levels are changing throughout development[12], it raises 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 and b), consistent with our observations from *in vitro* nucleosome reconstitution assays. Notably, in hindbrain we observed well-positioned nucleosomes at the center 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-value ≤ 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).

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* datasets, 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.

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 (Figure 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, though 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 and 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 and 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 be therefore 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 and see Supplementary Material for 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-value ≤ 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 (e.g. by the ε -amino group of lysine of histone proteins) to form a Schiff base (Fig. 4a).[14,15] To get 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 modeling to identify the critical, proximal lysine residues. Nucleosome core particles were assembled from 5fC-containing Widom 601 sequence then we employed 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 the irreversible trapping by chemical reduction may shift the equilibrium to the formation of Schiff base thereby potentially leading to the 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 (FDR value <0.0001, Benjamini-Hochberg correction on exact p-value from negative binomial distribution) were assessed by comparing the dataset 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 log2 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 other contexts, engaged in covalent Schiff base interactions between histones and DNA.

Next, we combined the data from polymerase stalling with molecular modeling 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 we quantified their contact frequencies (i.e. fraction of time a pair of atoms is less than 5 Å away; see Supplementary Material for details). The contact frequencies between bases and lysine side chains of histone proteins H2a/b, H3 and H4 are depicted in Fig. 4c (bottom panels) indicating that lysine residues from all four histone proteins are in close proximity to significant stalling sites. Notably, histone H3 displayed high contact frequency around 5fC sites within CpG context. Table 1 and Fig. 4d summarize our findings from polymerase stalling and molecular modeling, where predicted lysine residues are visualized in blue in the crystal structure that was used for the modeling with

proximal 5fC sites colored in red.

5fC can form a Schiff base with histones in chromatin

We next explored the existence of Schiff base interactions between 5fC of genomic DNA and lysines of histone proteins in the chromatin of mouse embryonic stem cells (mESC), to gain insights into the biological relevance of this interaction (Fig. 5a). We used $NaCNBH_3$ to chemically trap any Schiff base formed in nuclei from TDG KO mESC 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 qPCR analysis on synthetic DNA containing either 5fC or an abasic site with or without crosslinked lysine residue, confirmed that the polymerase was not able to efficiently amplify DNA with an abasic sites (with or without crosslinked lysine), whereas 5fC-containing DNA (with or without crosslinked lysine) was efficiently amplified (Supplementary Fig. 10).

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. 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-value = 7.7e-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) that 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-value < 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 analyzed 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 reduces the rate of RNA Pol II[29,28]. 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 ~2kb 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 caused strong promotion of nucleosome occupancy also 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

since 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 whereby 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 embryonic stem cells and mammalian tissues have identified patterns of 5fC that were distinct from 5hmC and 5mC.[32–34] These studies have shown that 5hmC is found enriched at promoters, gene bodies and poised enhancers, which is distinct from 5fC sites that are found 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. While a very recent report provided evidence that Schiff base interactions can form between 5fC and lysines in cells[14], that study did not identify proteins from which the key

lysine(s) originated. By chemically trapping 5fC-lysine conjugates before immunoprecipitation using antibodies against histone proteins, we now provide the first evidence that 5fC-histone(H3) conjugates can occur in mammalian cells. Notably our study revealed that in mESC 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.

Material and Methods

Chaperone assisted nucleosome assembly

Master mix was prepared following the manufacture's instruction (Chromatin Assembly Kit, Active Motif Belgium) with some modifications. Briefly, per 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 10X ATP Regeneration System. Complete 10X ATP

Regeneration System was prepared by mixing 0.1 µL Creatine Kinase with 1.65 µL 10X ATP Regeneration System. The mixture was gently agitated after addition of each component. DNA (100 ng) was diluted with ultrapure water to 7.5 µL, and mixed with 7.5 µL master mix to incubate at 27 °C (block) with 50 °C (lid) overnight. Finally, gel electrophoresis was performed for the quantification of the nucleosome fraction and free DNA.

In vitro reconstitution of chromatin

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

a) Hindbrain/ Heart genomic DNA reconstitution

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

b) 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). Per indexed library, cytosine or fully modified 5mC, 5hmC or 5fC DNA was subsequently generated by PCR (25 cycles) using the Taq polymerase and purified by GeneJET PCR Purification Kit and eluted with ultrapure water. 100 ng of each indexed libraries were 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 uL 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 uM, Sigma Aldrich), the supernatant was washed 2x through an Amicon Ultra spin column (10kDa cutoff, Merck). Input DNA libraries (before nucleosome reconstitution) and nucleosome fraction libraries were PCR amplified for 6 cycles with unmodified dNTPs and sequenced on 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 2x 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 2x 2 mL suspensions (1 control and 1 reduced sample) and repelleted for 15 min at 4°C (4000 rpm). 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 Covaris M220 using the truChIP Chromatin Shearing Kit (Covaris) following the manufactures guideline (at this point the size distribution of sonicated chromatin was assessed by tapestation after Genejet purification of a small aliquot). Protein denaturation was achieved using binding buffer containing a chaotropic agent (from Genejet PCR purification kit, Thermofisher) followed by 4x 450 µL water washes in the Amicon Ultra-Spin column (30kDa cutoff). After end repair, A-tailing and adapter ligation (no purification step), DNA libraries were used for chromatin immunoprecipitation (ChIP). For ChIP, 100 µL dynabeads M-280 sheep anti-rabbit IgG were washed and preincubated 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 ug salmon sperm DNA were added to the beads and incubated overnight at 4°C. Beads were then washed (5 min rotation at room temperature) 6x with 500 µL LiCl buffer (100 mM TRIS pH7.4, 500 mM LiCl, 1% NP40, 1% sodium deoxycholate) and 3x tube changes. For elution, beads were incubated (15 min rotation at room temperature) with 2x 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) 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

Methods used for the bioinformatics analysis are described in the Supplementary Material section.

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 the URL https://github.com/slab-bioinformatics/5fC-nucleosome .

Data availability

Sequencing data are available in the ArrayExpress database (www.ebi.ac.uk/arrayexpress) under accession number E-MTAB-6271.

5fC regions in tissues are available in the GEO database under accession number GSE77447. 5fC and 5caC regions in mESC are available under accession number GSE42250. Histone marks were obtained from ENCODE (https://www.encodeproject.org/) with the following accession numbers for hindbrain: H3K27ac (ENCFF203QTV), H3K4me1 (ENCFF542GAS) and for heart: H3K27ac (ENCFF954URD) and H3K4me1 (ENCFF737FNO). The datasets corresponding to the GRO-seq experiments were obtained from the GEO database under accession number GSE64748.

Author contribution

EAR, SB, WR and MAD designed the study. EAR, RH, ZL, WD, MI, JS and ZL performed the experiments. GP, SMC and DB performed the computational analysis. All authors analysed and interpreted the data. EAR, GP and SB wrote the manuscript, with contributions from all authors.

Ethics approval and consent to participate

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

Competing interests statement

SB is a founder, advisor and shareholder of Cambridge Epigenetix Ltd.

Figure legends

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-labeled DNA was used for quantification. Non-denaturing polyacrylamide gel (4-16%) electrophoresis was run in 0.4x TBE buffer to separate the nucleosome fraction from the free DNA. Experiments were repeated independently with similar results (n = 3). **(b)** Log2FC of nucleosome occupancy (nucleosome/total DNA) of modified DNA compared to unmodified Widom sequence. The error bars represent the standard deviation of the mean from three independent

experiments (*n* = 3). Two-tailed t-test was used to calculate p-values (5mC p-value = 0.18, 5hmC p-value=3.3e-3 and 5fC p-value = 6.8e-3). **(c)** Nucleosomes were reconstituted using different 5fC densities (100% = fully modified to 1% ~ 1 5fC per sequence). Normalized nucleosome fractions were plotted against 5fC density. The error bars represent the standard deviation 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 5fC-modified Widom sequence and are represented as the mean along with the standard deviation of the mean from three independent experiments (*n* = 3). Two-tailed t test was used to calculate p values (* p = 0.0363, p-value= 0.0113 and p-value= 0.0123, ** p-value = 0.0039). **(e)** A pool of DNA sequences containing either cytosine, 5mC, 5hmC or 5fC was formed by combining sub-libraries 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* = 2977), 5mC (orange, *n* = 884), 5hmC (green, *n* = 10537) and 5fC (red, *n* = 2481) DNA was plotted. Notched boxplot shows the 1st, 2nd (median) and 3rd quartile, with whiskers extending to the minimum and maximum. P-values were obtained using the two-sided Mann–Whitney test (**** p-value ≤ 0.0001, C vs 5mC p-value=6e-9, C vs 5hmC p-value=4.8e-16, C vs 5fC p-value 7e-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 RPKM. Data has been normalized to library input. Experiments were repeated twice independently with similar results (*n* = 2).

Fig. 2. 5fC within a genomic sequence context enhances nucleosome occupancy in vitro

(a) and **(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 shades in Fig. 2a shows the standard error of the mean 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 (i.e. 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 4000 bp regions 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).

***Fig. 3. 5fC is a determinant of nucleosome organization in vivo* that is linked to gene expression**

(a) and **(b)** Normalized MNase signal (RPKM) was plotted around 5fC sites (dotted black line) in hindbrain (blue) and heart (orange) revealing increased nucleosome density at the center of 5fC sites. **(c)** Notched boxplot shows the nucleosome occupancy in hindbrain and heart tissues. Boxplot shows 1st, 2nd (median) and 3rd quartile, with notches representing the confidence interval around the median. Average nucleosome occupancies within 5fC sites in heart (*n* = 11226, orange) and hindbrain (*n* = 16732, blue) are significantly higher (p-value = 5.8e-17 for heart, p-value < 3.3e-306 for hindbrain [bellow machine precision], two-sided Mann-Whitney U-test) than that of all detected nucleosomes in heart (*n* = 10936694) and hindbrain (*n* = 10973437). **(d)** Venn diagram showing the overlap of nucleosomes at heart- and hindbrain-unique 5fC sites. **(e)** and **(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* = 13286, dark blue) in hindbrain and 5fC-nucleosomes (*n* = 827, orange) and all enhancer sites (*n* = 13286, red) in heart. Notched boxplot show that the presence of 5fC at enhancer sites correlates with significantly higher gene expression (p-value = 3.0e-8 for hindbrain, p-value=2.2e-4 for heart, two-sided Mann-Whitney U-test) of their associated genes compared to all expressed genes (predicted enhancer-gene list from Shen et al[23]). Boxplot shows 1st, 2nd

(median) and 3rd quartile, with notches representing the confidence interval around the median and the black diamond the mean, and the whiskers indicate the reach of the data points beyond the 1st (Q1) and 3rd (Q3) quartile (e.g. Q1+1.5*(Q3-Q1)).

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)** Upper panel: Log_2 fold change between the number of reads of crosslinked and control sample for the forward (blue) and reverse strand (orange). The gray sinusoidal line indicates the orientation of the major groove with respect to the histone core, ranging from 1 (histone-core facing) to -1 (facing away from the histone core). Significant (FDR values of 9.6e-5, 9.6e-5, 1.1e-4 and 9.6e-5 from left to right, Benjamini-Hochberg correction on exact p-value from negative binomial distribution) stalling sites (±3 bases) around CpG dinucleotides within the DNA sequence are highlighted in the DNA sequence in grey. Lower panels: 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 the Schiff base formation were identified. Experiments were repeated twice independently with similar results ($n = 2$). **(d)** Structure highlights 5fC sites (red balls) within the DNA and lysine residues (blue balls) that potentially interact through the formation of Schiff-base conjugates.

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

(a) Workflow for the detection of *in vivo* Schiff base sites in mESC TDG KO. 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)** Log2FC of reduced/untreated samples after ChIP-qPCR reveal significant enrichment (two-tailed t-test, p-value= 7.7e-3) of H3 immunoprecipitated chromatin after reduction. 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 standard deviation. 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 mESC TDG KO. 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) centered around 25 kb up- and downstream of Schiff base 5fC-H3 sites (blue) and 5fC only (red) sites.

Table 1

Table gives details on crosslinked 5fC position, 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

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,** 1323–9 (2014).

5. Jimenez-Useche, I. *et al.* DNA Methylation Regulated Nucleosome Dynamics. *Sci. Rep.* **3,** (2013).

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 (80-.).* **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.* (2017). doi:10.1038/nrg.2017.33

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. Chemie 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.* **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.* **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 (80-.).* **322,** 1845–1848 (2008).

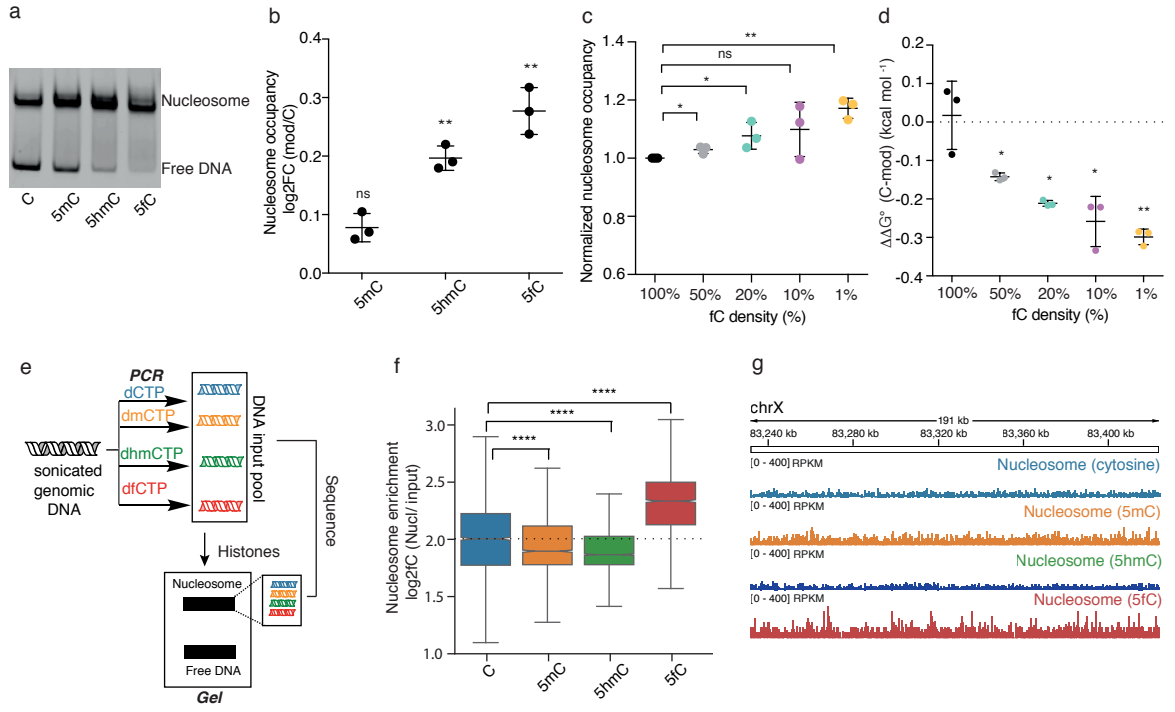


Figure 1

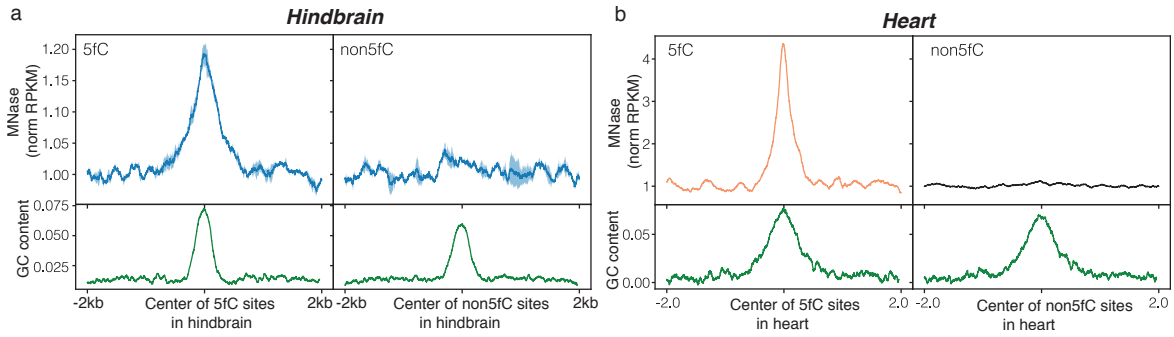


Figure 2

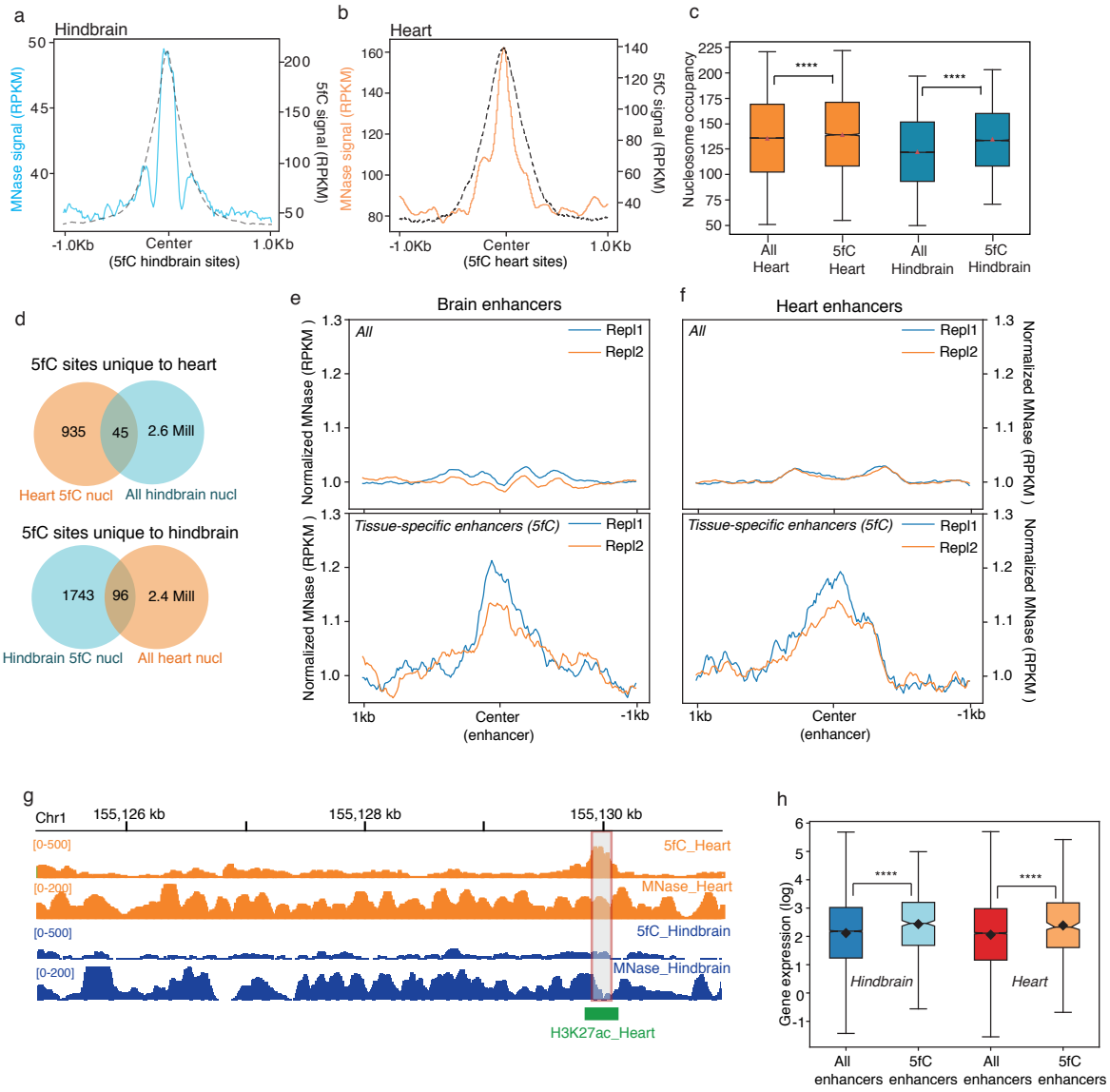

a
Hindbrain
MNase signal (RPKM)
5fC signal (RPKM)
-1.0Kb
Center
1.0Kb
(5fC hindbrain sites)
b
Heart
MNase signal (RPKM)
5fC signal (RPKM)
-1.0Kb
Center
1.0Kb
(5fC heart sites)
c
Nucleosome occupancy
****
****
All Heart
5fC Heart
All Hindbrain
5fC Hindbrain
d
5fC sites unique to heart
935
45
2.6 Mill
Heart 5fC nucl
All hindbrain nucl
5fC sites unique to hindbrain
1743
96
2.4 Mill
Hindbrain 5fC nucl
All heart nucl
e
Brain enhancers
All
Repl1
Repl2
Normalized MNase (RPKM)
Tissue-specific enhancers (5fC)
1kb
Center
(enhancer)
-1kb
f
Heart enhancers
All
Repl1
Repl2
Normalized MNase (RPKM)
Tissue-specific enhancers (5fC)
1kb
Center
(enhancer)
-1kb
g
Chr1
155,126 kb
155,128 kb
155,130 kb
[0-500]
5fC_Heart
[0-200]
MNase_Heart
[0-500]
5fC_Hindbrain
[0-200]
MNase_Hindbrain
H3K27ac_Heart
h
Gene expression (log)
****
****
Hindbrain
Heart
All enhancers
5fC enhancers
All enhancers
5fC enhancers

Figure 3

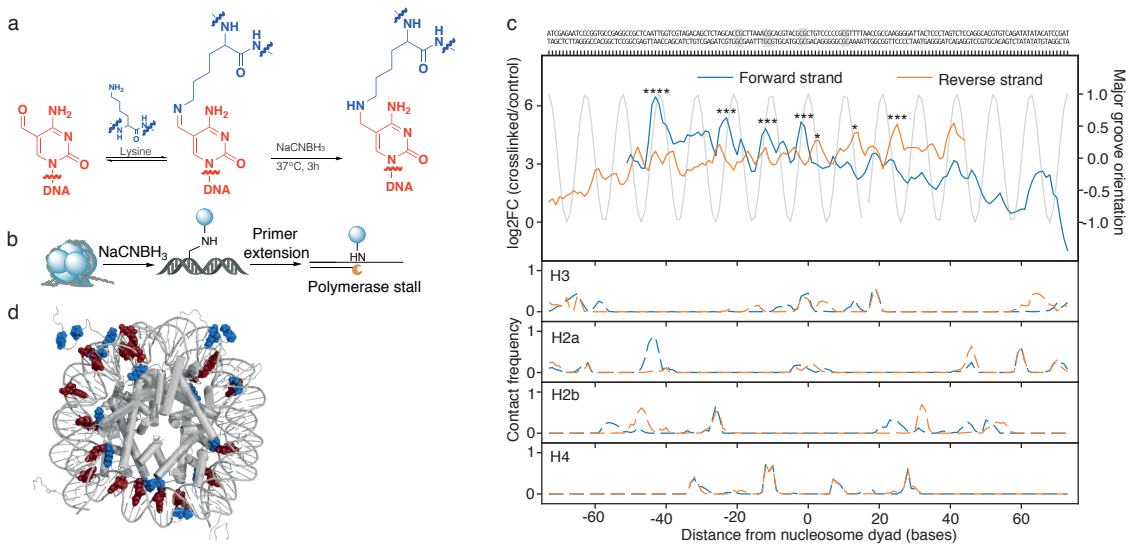

a
Lysine
NaCNBH3
37°C, 3h
DNA
b
NaCNBH3
Primer
extension
Polymerase stall
c
Forward strand
Reverse strand
log2FC (crosslinked/control)
Major groove orientation
Contact frequency
H3
H2a
H2b
H4
Distance from nucleosome dyad (bases)
d

Figure 4

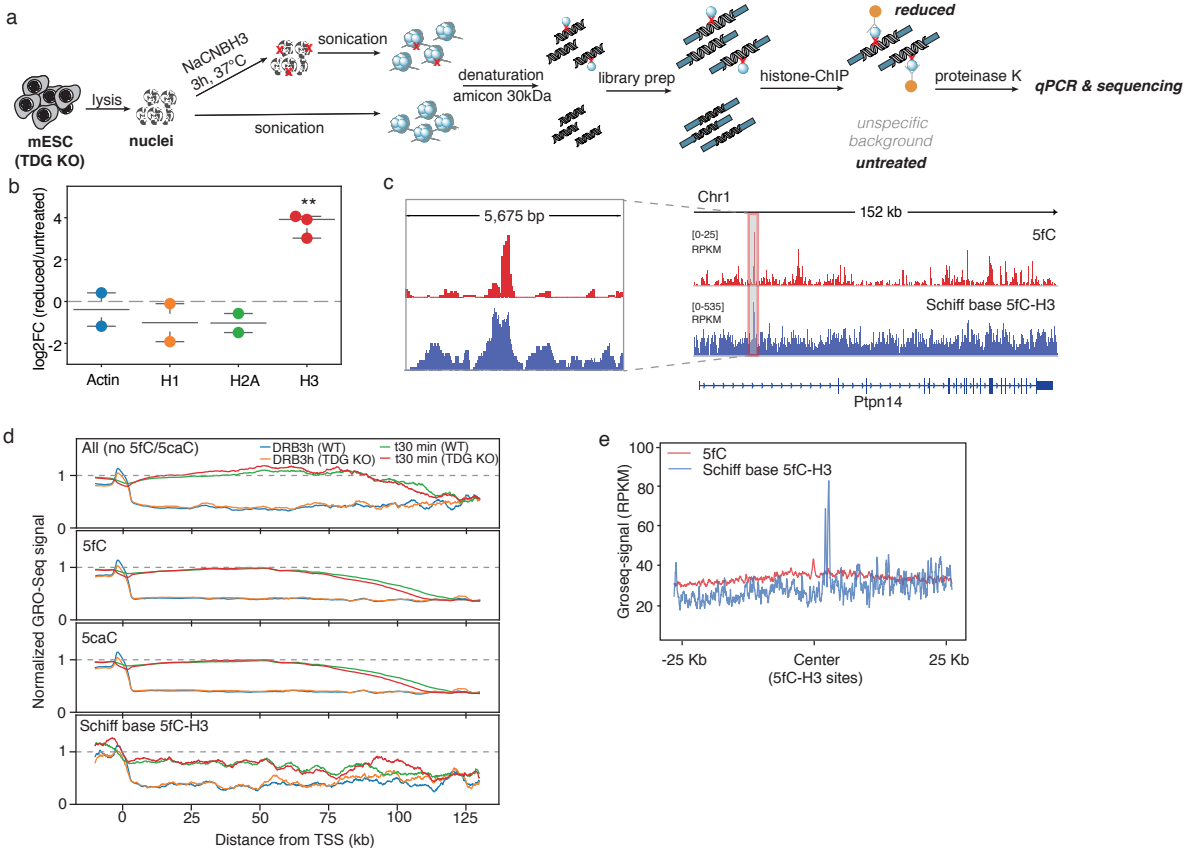

a
mESC
(TDG KO)
lysis
nuclei
NaCNBH3
3h, 37°C
sonication
sonication
denaturation
amicon 30kDa
library prep
histone-ChIP
reduced
unspecific
background
untreated
proteinase K
qPCR & sequencing
b
log2FC (reduced/untreated)
**
Actin
H1
H2A
H3
c
5,675 bp
Chr1
152 kb
[0-25]
RPKM
5fC
[0-535]
RPKM
Schiff base 5fC-H3
Ptpn14
d
All (no 5fC/5caC)
DRB3h (WT)
DRB3h (TDG KO)
t30 min (WT)
t30 min (TDG KO)
5fC
5caC
Schiff base 5fC-H3
Normalized GRO-Seq signal
Distance from TSS (kb)
e
5fC
Schiff base 5fC-H3
Groseq-signal (RPKM)
-25 Kb
Center
(5fC-H3 sites)
25 Kb

Figure 5

SUPPLEMENTARY INFORMATION

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

Authors: Eun-Ang Raiber, Guillem Portella, Sergio Martínez Cuesta, Robyn Hardisty, Pierre Murat, Zhe Li, Mario Iurlaro, Wendy Dean, Julia Spindel, Dario Beraldi, Zheng Liu, Mark A. Dawson, Wolf Reik, Shankar Balasubramanian

INDEX

Table of Contents

Supplementary Methods

1.1 DNA Preparation

DNA for nucleosome assemblies was generated by PCR amplification with Taq polymerase (NEB, Massachusetts, USA) and using unmodified dNTPs (Thermo Scientific) and/or modified deoxycytidine triphosphates (TriLink California, USA) and cy3 or cy5 labeled primers (Biomers, Germany). Low-density modified DNA sequences were generated by PCR using a pool of modified xCtp and dCtp (see Supplementary Table 3 for ratios). PCR products were purified by GeneJET PCR Purification Kit (Life Technologies, California, USA) and eluted with ultrapure water.

1.2 Competitive nucleosome reconstitution by dialysis (free energy measurements)

Nucleosome assembly was prepared by dialysis (histones were obtained from the *EpiMark® Nucleosome Assembly Kit, NEB*) in the presence of 4.2 µg competitor DNA (5SrDNA obtained from NEB). Briefly, cy3-labelled Widom DNA (100 ng) in 2 M NaCl was mixed with 125 nM histone dimer and tetramer and competitor DNA in a total reaction volume of 50 uL. After incubation at 4°C for 2 h, the mixture was sequentially dialysed into 20 mM TRIS, pH8 buffer containing 1 mM EDTA, 1 mM DTT and 1.5 M, 1 M, 0.6 M or 0.25 M NaCl. Each dialysis step lasted for at least 2h, with the dialysis into 0.6 M NaCl being overnight. Finally gel electrophoresis was performed for the quantification of the nucleosome fraction and free DNA. Two technical replicates were generated for each modification. For the calculation of the free energy ($\Delta G°$), K_{eq} was first determined as the ratio of the nucleosome fraction to free DNA (Cy3-labeled nucleosome fraction/ Cy3-labeled free DNA). The corresponding free energy was then obtained from $\Delta G° = -RT\ln(K_{eq})$, with RT = 0.55 kcal mol^{-1} (T = 4°C). Difference free energies

($\Delta\Delta G°$) were calculated for each sample as: $\Delta\Delta G°_{modification} = \Delta G°_{modification} - \Delta G°_{cytosine}$.

1.3 Gel electrophoresis

Samples were run on 6% DNA Retardation gel (6%, Thermo Scientific) in 0.4XTBE buffer at 4°C at 200 V using XCell SureLock Mini-Cell Electrophoresis tank (Thermo Scientific). Gels were pre-ran for at least 1 h. Gels were imaged with Typhoon Trio Imager (Amersham Biosciences, UK) with 532 nm excitation and 580 emissions for Cy3, 633 nm excitation and 610 nm emissions for Cy5. The band intensity was quantified using ImageJ.

1.4 Tissue dissection

Breeding and tissue collection was performed as previously described (Iurlaro et al 2016[1]).

1.5 MNase sequencing of reconstituted nucleosomes

After incubation at 27°C for overnight, 0.1 M $CaCl_2$ (3 μL) was added together with 1 μL of the enzymatic shearing cocktail and incubated for 4 min at room temperature. 35 μL of the 4x enzymatic stop solution was added and the mixture was incubated on ice for 10 min before 20% SDS (final concentration of 0.5%) was added. After proteinase K digestion, the DNA was obtained by standard chloroform/phenol extraction and ethanol precipitation. After DNA library preparation libraries were pooled and sequenced on the Illumina NextSeq 500. For the analysis of the MNase datasets, replicate experiments obtained from hindbrain DNA reconstitution were analysed separately and results were averaged. Datasets obtained from replicate experiments (heart) were merged and subsequently analysed due to low sequencing depth of replicate 1.

1.6 Native MNase sequencing on tissues

Tissues obtained from mouse embryos were homogenized, washed with 5mL cold 1xPBS and pelleted at 300g at 4 °C for 10 min. Cell pellet was resuspended in 500 μL of ice-cold NP40 lysis buffer (10 mM Tris-HCl pH7.4, 10 mM NaCl, 3 mM $MgCl_2$, 0.5% Nonidet P-40, 0.15 mM spermine and 0.5 mM spermidine) and incubated on ice for 5 min. Nuclei were subsequently pelleted at 120g at 4 °C for 10 min, washed with MNase digestion buffer (10 mM Tris-HCl pH 7.4, 15 mM NaCl, 60 mM KCl, 0.15 mM spermine and 0.5 mM spermidine) and resuspended in MNase digestion buffer containing 1 mM $CaCl_2$. MNase (5 units) were added and the sample was incubated at room temperature for 5 min before the addition of MNase digestion buffer and MNase stop buffer (100 mM EDTA and 10 mM EGTA pH 7.5). Proteinase K and 20% SDS was added and the mixture was incubated at 37 °C overnight. After chloroform/ phenol extraction, 2 μL of heat-treated RNase A (10 mg/ mL) was added to each sample and incubated at 37 °C for 2 h. After another round of chloroform/ phenol extraction, the digested genomic DNA was obtained by ethanol precipitation. Two biological replicates per tissue from WT and TDG KO were generated. For the analysis of the MNase datasets, replicate experiments were analysed separately and the results combined.

1.7 *In vitro* Schiff' base reduction for gel analysis

Nucleosomes were treated with 100mM $NaBH_3CN$ and incubated at 37 °C for 18 h. The mixture was desalted and purified by micro-biospin P6 Tris columns (BioRad), and subsequently run on a 12% Bis-Tris Protein gel.

1.8 Polymerase stop assay

Reduced nucleosomes using either unmodified or 5fC (100%) modified Widom 601 sequence were obtained as described above. Polymerase extension was achieved in the presence of dNTP (200 μM), primers (0.5 μM), 1 x polymerase buffer and DreamTaq (1 μL, 5U). The mixture was heated at

95 °C for 30 s, 60 °C for 60 s and 72° C for 3 min. After extension, the mixture was treated with Proteinase K (40 μg) in Proteinase K buffer (750 mM Gu-HCl, 5% Tween 20, 30mM EDTA, 30mM Tris-HCl), before purification by Oligo clean and concentrator Kit (Zymo Research). The samples were treated with RecJF (3 μL, 90U) for 12 h at 37 °C to remove excess primer and overhangs. The samples were again purified using the Oligo clean and concentrator Kit before DNA library preparation for sequencing. Two technical replicates were generated for each condition, control DNA (free DNA) and "cross-linked" DNA.

1.9 RNA sequencing (hindbrain and heart tissues)

Total RNA for each sample was extracted using RNeasy mini kit (QIAGEN) and following the manufacturer's instructions. Extracted RNA was polyA-enriched. Then, RNA was used for library preparation using the TruSeq RNA Library Prep Kit v2 (Illumina) following the manufacturer's instructions. Libraries were indexed using Illumina Indexes and 100bp paired-end sequencing was performed on Illumina HiSeq 2000 instrument using TruSeq reagents (Illumina), according to manufacturer's instructions. Three biological replicates per tissue were generated from WT and TDG KO.

1.10 Cell culture

TDG -/- mouse embryonic stem cells[2] were cultured on gelatinised plates in DMEM medium supplemented with 15% fetal bovine serum, 1% Penicillin Streptomycin, 1% MEM non-essential amino acids, 1% Glutamax, 50 µm β-mercaptoethanol and 10^3 U LIF, in normoxic conditions at 37°C.

1.11 Molecular Dynamics simulations of nucleosome particles

Molecular dynamics (MD) simulations were carried out using the Gromacs-4.5 software[3], using periodic boundary conditions and the particle mesh Ewald method[4] for the long-range electrostatics. The short-range repulsive and

attractive dispersion interactions were modeled via a Lennard-Jones potential with a cut-off of 1.0 nm. The Settle algorithm[5] was used to constrain bond lengths and angles of water molecules, and P-Lincs[6] was used for all other bond lengths, in combination with virtual interaction-sites[3,7] to remove the hydrogen vibrations and therefore use a time step of 4 fs. The temperature was kept constant as described in Bussi et al.[8]. The pressure was kept constant and it was controlled by coupling the simulation box to a pressure bath of 1 atm[9]. The amber99SB*-ildn[10] force-fields was used to describe the histone tails, and the amber99+parmBSC0[11] force field was used for the nucleosome DNA. The solvent was modeled using the TIP3P water model, the sodium and chlorine ion were modeled using Dang's parameters[12], and manganese atoms parameters taken from the Amber force field database.

Two dinucleotide model initial conformations were built following the protocol described in Collepardo et al.[13], by stacking two nucleosome particles (X-ray structure with PDB code 1KX5)[14] on top of each other. In one model the inter-nucleosome distance, as measured by the vector connecting the center of mass of the two nucleosomes was set to 6 nm, and in the second model was set to 7 nm. The histone tail sequences were replaced by the human sequences. The di-nucleosome systems were embedded in a truncated octahedron box containing ~200,000 water molecules, leaving 2 nm between the nucleosome atoms and the edges of the box. This separation is large enough to accommodate a fully extended H3 tail, which is the longest one. Approximately 900 sodium ions and 600 chlorine ions were added to balance the nucleosome charge and give an ionic concentration of 150 mM NaCl (the exact values depend on the model). Each di-nucleosome system was energy minimized and simulated twice (using two different random seeds) for 1.15 microsecond.

1.11.1 Map of intranucleosome DNA – lysine contacts

To facilitate the analysis the MD trajectories of the di-nucleotide systems were split into individual nucleosomes, imaged to remove periodic boundary crossings, and then concatenated into one long trajectory containing one nucleosome. To alleviate auto-correlation effects, we analysed frames with a 2 ns frequency. After

discarding first 100 ns of each trajectory, the resulting concatenated trajectory contained 4200 structures. For each nucleotide, we collect the set of lysine residues whose side chain atoms are found within a cut-off distance of 1.2 nm with respect to the nucleotides in any of the analysed frames. From these set of distances, we compute for each pair of reference position – lysine side-chain a time/ensemble averaged contact metric by means of a continuous switching function. We have used $1/\left(1+e^{\left(b*(x-1.5*d0)\right)}\right)$ as such switching function, were b = 10 and d0 = 0.5 and x represents the minimum distance between any atom in the lysine side chain with the reference nucleotide. The parameters were empirically chosen such that distances below 0.5 nm result in a value of ~1, and anything above 0.5 nm decays to 0 (at ~1 nm is almost zero).

1.11. 2 Groove orientation

As proxy for the orientation of duplex DNA strands with respect to the nucleosome core we defined as phase angle φ the angle between a v_{bp} vector centered at the base pair center of mass and pointing towards the minor groove with the vector connecting the center of mass of the base pair and the center of mass of the nucleosome. The vector v_{bp} was defined as the sum of the two vectors connecting the center of mass with the N9(Y)/N1(R) for a given base pair. The resulting curve was further refined by fitting a sinusoidal curve, and we rescaled the maximum and minimum values to the [-1, 1] range for visualization purposes. In our definition, a large φ value associated to a given base pair reports locations where the major groove faces the histone core.

1.11.3 Integrating polymerase stop assay signals with MD simulations

We combined the outcome of the polymerase stop assays signals with a MD-derived probability map of lysine-nucleotide contacts to produce a set of high-probability locations 5fC-lysine cross links. The two sequence-dependent signals were multiplied to create a mixed signal, which is high in locations were both metrics indicate high-probability of cross-linking. Peaks were detected using derivatives of the combined metric, employing the PeaksUtils 1.1 python package

with an amplitude threshold of 0.25 and a minimum distance of 1 base pair between peaks. For each peak, we reported the closest C residue, and it's associated lysine residues.

1.12 qPCR analysis of polymerase activity over abasic sites

Abasic DNA was generated by treatment of abasic-ODN with uracil-DNA glycosylase (10 U, NEB) at 37 °C for 2 h, then purified using an oligo clean & concentrator kit (Zymo Research). Abasic and 5fC control DNA (200 ng each) were incubated with lysine (500 mM) and $NaBH_3CN$ (100 mM) in PBS at 37 °C for 3 h. As a control, the same input DNA was used without lysine and $NaBH_3CN$ treatment. Samples were desalted using a micro bio-spin P6 column and diluted 1000-fold before qPCR amplification on a CFX96 Touch Real-Time PCR Detection System. Reactions (10 µL) contained Q5 High-Fidelity 2X Master Mix (5 µL), primers (1 µM), SYBR Green stain (0.3X) and diluted DNA (1 µL). The Ct value obtained for each sample was compared to that of the untreated 5fC ODN.

1.13 Bioinformatics and data analysis

Reads in fastq format obtained from the Illumina sequencing pipeline have been trimmed to remove the sequencing adapters and aligned against the mouse genome (NCBI version mm9) using bwa[15] with default settings. Subsequently, we removed unmapped, not primarily aligned, supplementary aligned, low quality and optical duplicate reads. We calculated the coverage of non-duplicate and extended reads across the genome by means of bamCoverage[16] with a bin size of 1 base pair, normalized to RPKM (reads per kilobase per million). We used deeptools compute-matrix[16] to calculate the average genomic coverage centered around the set of common 5fC sites across duplicates within a region of ± 2000 bp, both for heart and for hindbrain. As control, we calculated the average genomic coverage on the same number of non-5fC sites (i.e. 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 4000 bp regions matched that of the 5fC regions with a root-mean square error below 0.0025.

1.13.1 Peak calling

To detect regions significantly enriched in reads we used MACS2[17] peak calling in narrow peak mode using with default settings for pair-end reads.

1.13.2 Nucleosome positioning and occupancy

We used iNPS[18] to identify nucleosome positioning in each tissue (hindbrain and heart) from our MNase-seq data sets. After visualizing the read-size distribution of the pair-end MNase-seq experiment, we used a pair-end maximum of 230 bases (iNPS option *--pe_max=230*) and a pair-end mimimum of 100 bases (iNPS option *--pe_min=100).*We used DANPOS[19] Dpos module with default settings to calculate the differential nucleosome occupancy between different tissues.

1.13.3 5fC-associated nucleosomes

We used nucleosome positions identified by iNPS to overlap the nucleosome map with 5fC peaks. "5fC-associated nucleosomes" were called when nucleosome positions overlapped by at least 80% with 5fC peaks.

1.13.4 Differential gene expression

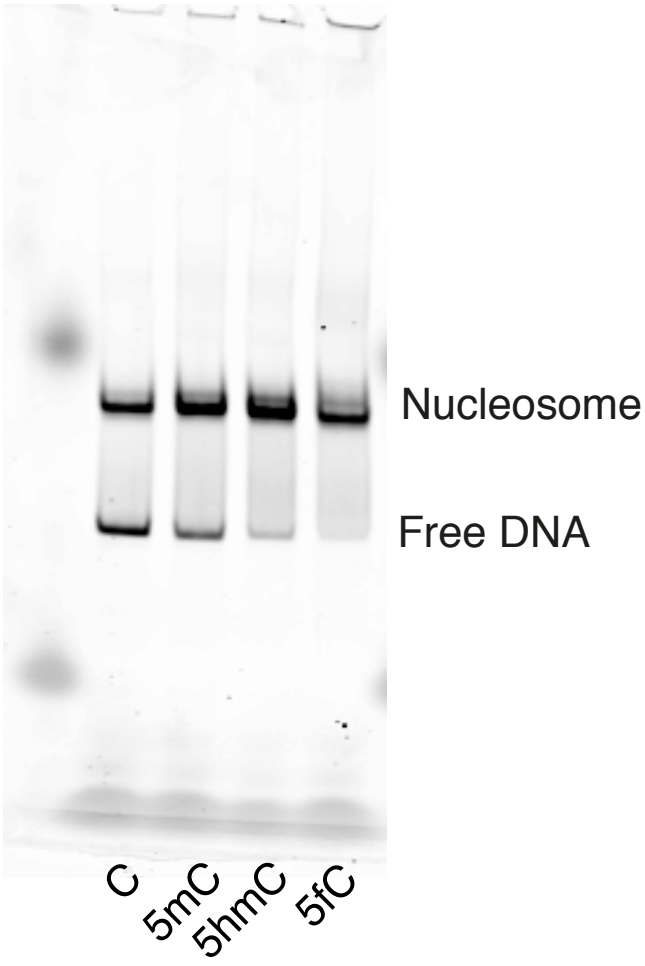
We used Trim-Galore to remove the adapters and filter out low quality reads (options -q 10 --stringency 8). The resulting reads were aligned to the reference mm9 genome using Hisat2[20], using Mus_musculus.NCBIM37.67.gtf to generate a list of splice-sites for the alignment. After alignment, the different sequencing lanes of the same library were merged together, and gene counts were generated via Htseq-counts[21]. Differential expression was computed using the generalized linear model method implemented in the R package edgeR[22].

1.13.5 Changes in time dependent Pol II transcription elongation in the presence of 5fC

We assessed the potential impact of endogenous 5fC-histone interactions on the rates of Pol II transcription elongation in TDG KO mESCs using previously publicly available data on global nuclear run-on coupled with deep sequencing (GRO-seq)[23]. We used three different sequencing sets (with duplicates) corresponding to different time points in WT and TDG KO: no DRB treatment (NODRB, equivalent to steady state), 3 h of DRB treatment (DRB3h) and 30 min after washing away DRB (t30 min). Sequencing reads for each condition and time point were obtained from GEO GSE64748. For each data set, we removed sequencing adapters and poly-A tails in each read, and retained only reads longer than 16 base pairs. We aligned the resulting reads with *bwa* for single end reads against the mouse genome (NCBI mm9), allowing a maximum of 2 mismatches, and we filtered out optical duplicates.

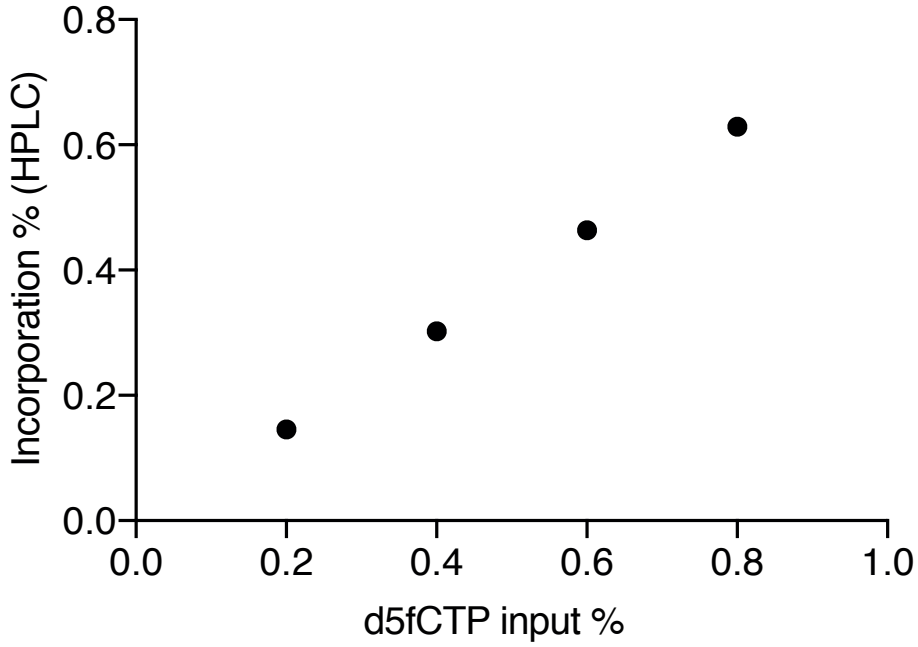
We determined the location of the Pol II transcription wave at each time point by means of metagene profiles, closely following the methodology described in Wang et al Nature [23]. We performed the analysis both at the transcript and gene level, and since the results were qualitatively equivalent, we present here the analysis at transcript level. We retained for the analysis only those transcripts, which had more than 0.5 RPKMs in the steady state (NODRB) sample. The transcripts were all aligned at their TSS, and we averaged the read coverage, normalized to RPKM, for all active transcripts using 100 base pair bins. The resulting profiles were normalized using the steady state curve, and the final curves were smoothed using a Savitzky-Golay filter (5.1 kb window, order 1 polynomial). Metagene profiles for transcripts falling into different subsets (e.g. Crosslinked 5fC-H3 sites) were generated by the same procedure using only the set of transcripts, which intersected with GRO-seq active transcripts. The genomic location of 5fC and 5caC in mESC were obtained from the datasets of Shen et al.[24]. We restricted our metagene analysis on the set of in vivo crosslinked 5fC sites to those that intersected with 5fC sites, as well as to the active transcripts.

Supplementary Figures



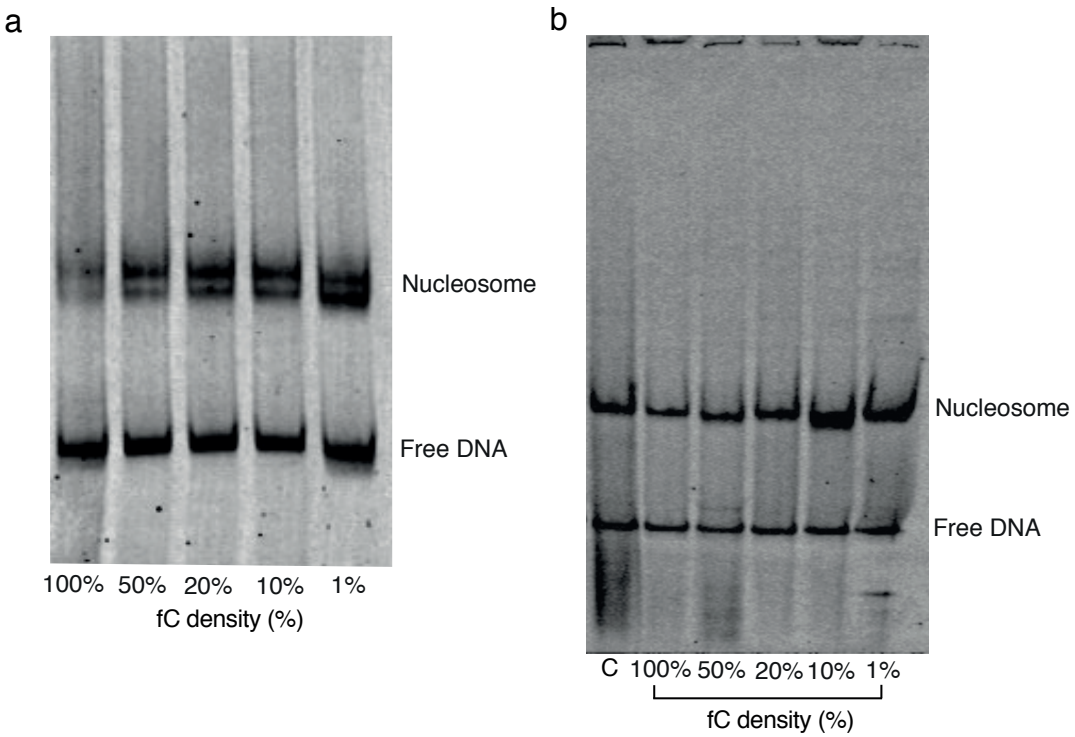
Supplementary Fig. 1

Full image of the non-denaturing polyacrylamide gel shows the nucleosome and free DNA fraction of cytosine, 5mC, 5hmC and 5fC-modified Widom. Experiments were repeated independently with similar results ($n = 3$)



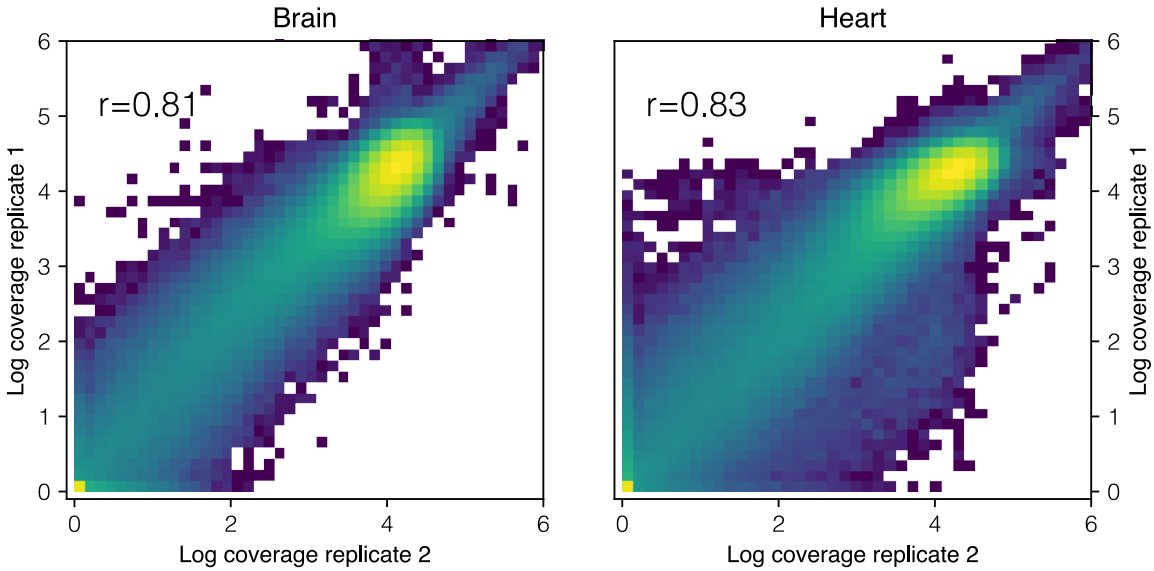
Supplementary Fig. 2

HPLC analysis of nucleotides after digestion was used to measure the actual incorporation (%) of d5fCTP during PCR to generate different 5fC density containing Widom DNA.



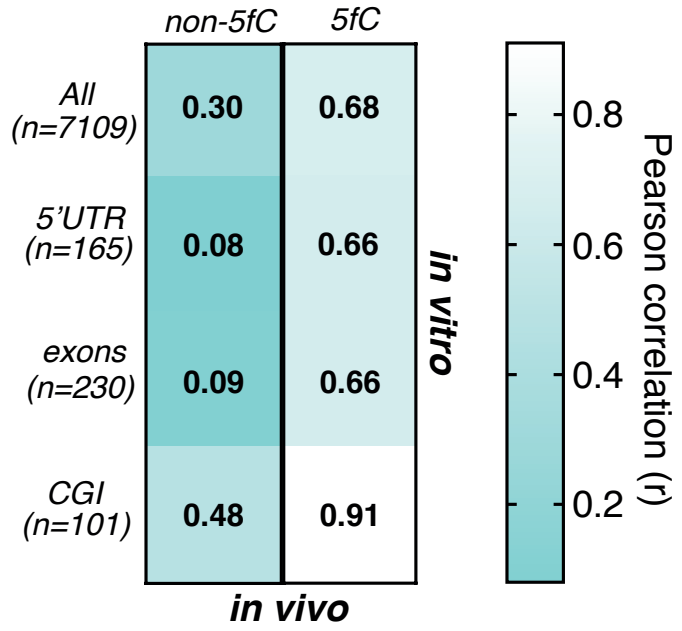
Supplementary Fig. 3

(a) Image of the non-denaturing polyacrylamide gel shows the nucleosome and free DNA fraction when different 5fC-density containing Widom sequence was used for nucleosome assembly (chaperone assisted) **(b)** Quantification of the nucleosome and free DNA fraction after nucleosome assembly by dialysis was used for the calculation of free energies. Experiments were repeated independently with similar results ($n = 3$)



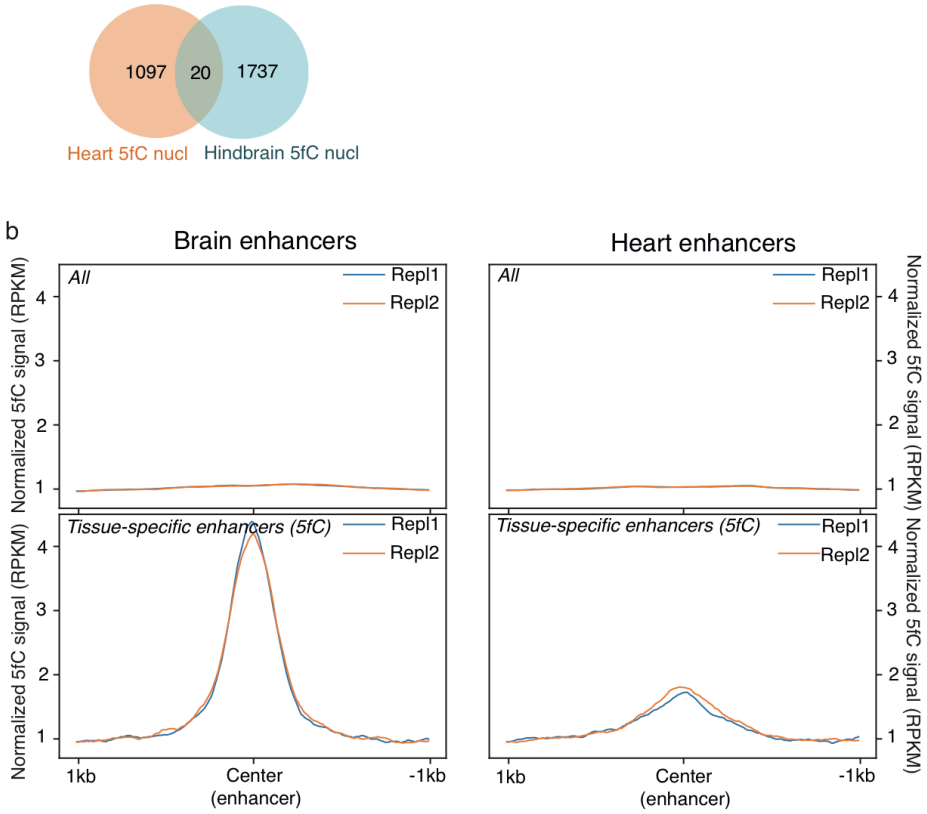
Supplementary Fig. 4

Scatter plots obtained by comparing MNase-seq datasets from two biological replicates (hindbrain and heart) show good reproducibility. For each condition, we have divide the genome in 1 kb non-overlapping bins, except the last bin for each chromosome which is smaller, resulting in a sample size of 2472370. We quantified the read coverage (in RPKM) in those regions and calculated the correlation between the coverage using the Pearson correlation coefficient.



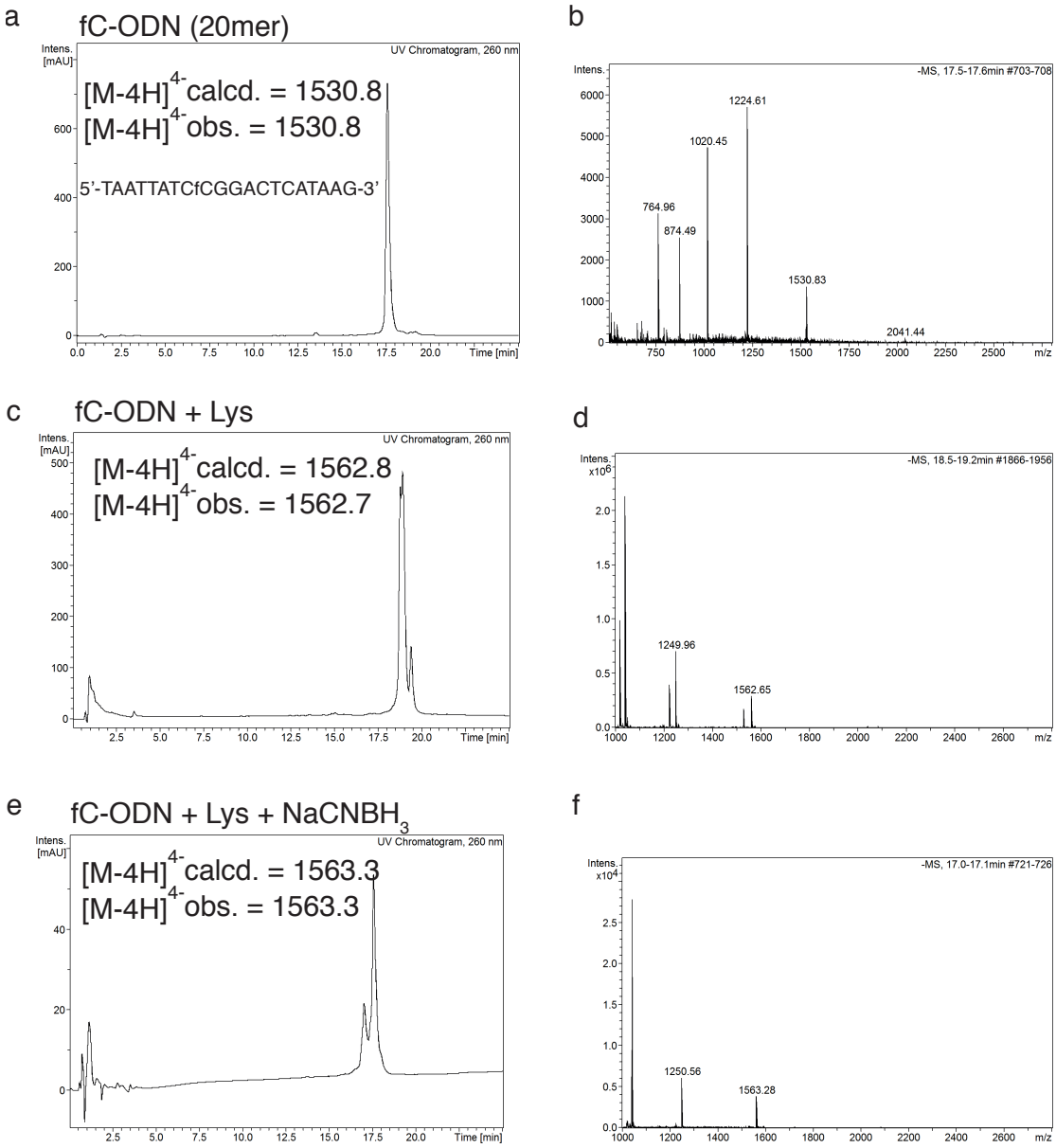
Supplementary Fig. 5

Comparison of *in vitro* MNase signals generated by nucleosome reconstitution with *in vivo* MNase signals, displaying higher Pearson correlations of MNase signals at 5fC sites compared to non5fC sites. This observation demonstrates an intrinsic 5fC-DNA preference of nucleosomes supporting a role for 5fC in determining the organization of nucleosomes.



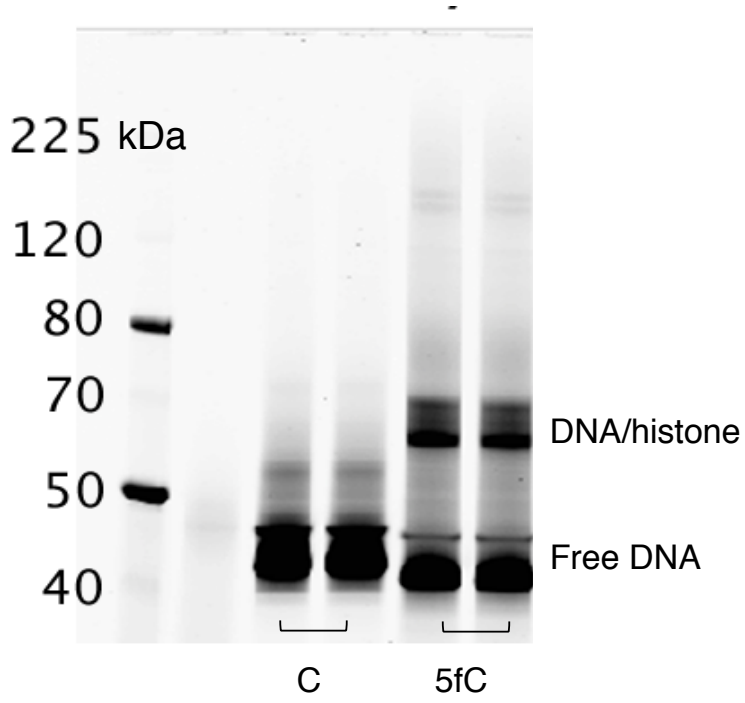
Supplementary Fig. 6

(a) Venn diagram showing the overlap of nucleosomes at heart- and hindbrain- 5fC sites. **(b)** Tissue-specific enhancers are enriched in 5fC as measured by 5fC signals (RPKM) around the center of enhancers.



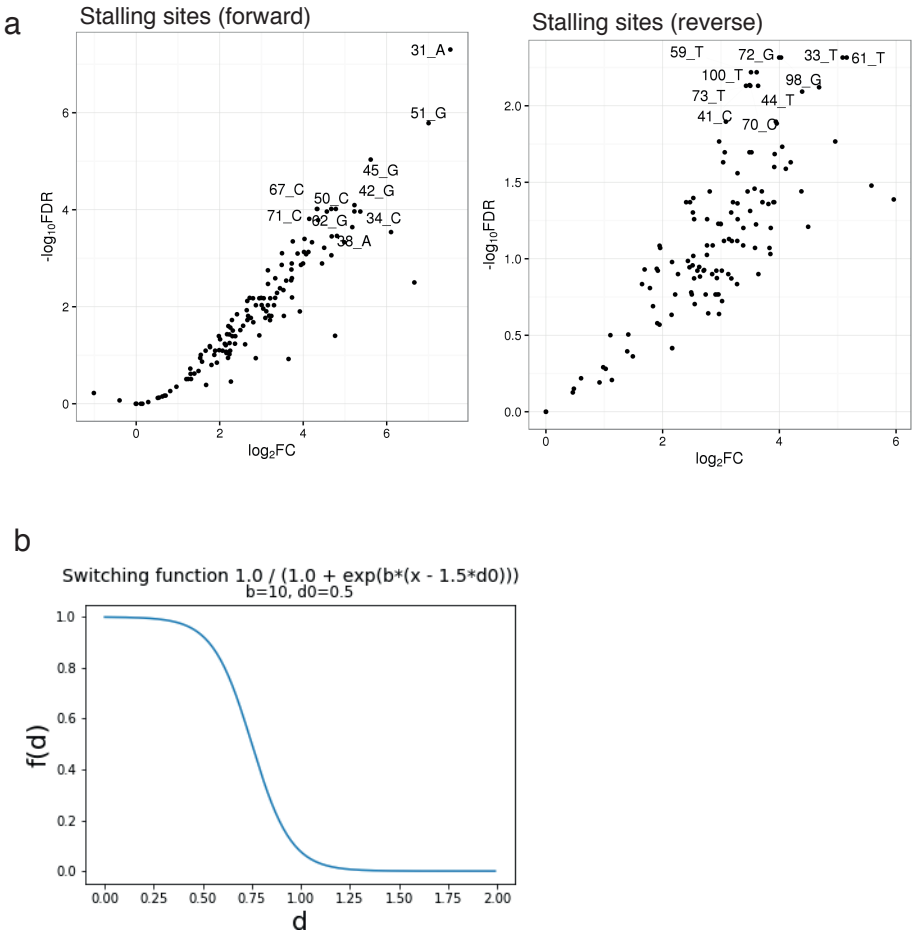
Supplementary Fig. 7

LCMS traces are shown for 5fC-ODN **(a,b)**, 5fC-ODN+Lys **(c,d)** and reduced 5fC-ODN + Lys **(e,f)**. Experiments were repeated twice independently with similar results ($n = 2$).



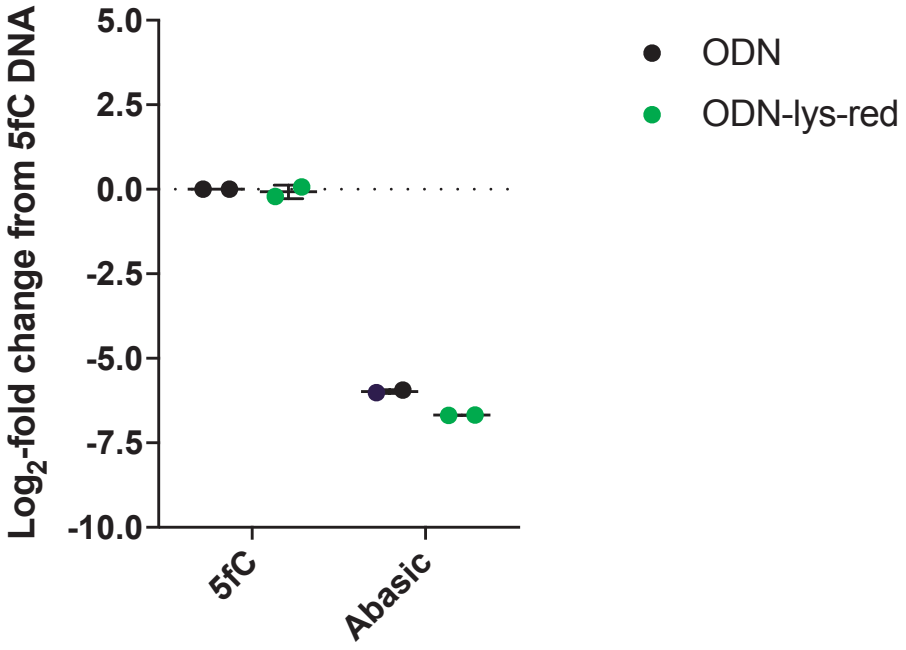
Supplementary Fig. 8

The nucleosome preparations with 5fC and C (control) were, after $NaBH_3CN$ treatment, resolved by 4-12% SDS-PAGE denaturing gel electrophoresis to separate non-crosslinked DNA. Two replicates (n=2) are shown for each, cytosine and 5fC (both cy3-labeled). The appearance of a second, higher DNA band was observed for 5fC DNA due to covalent DNA-protein complex.



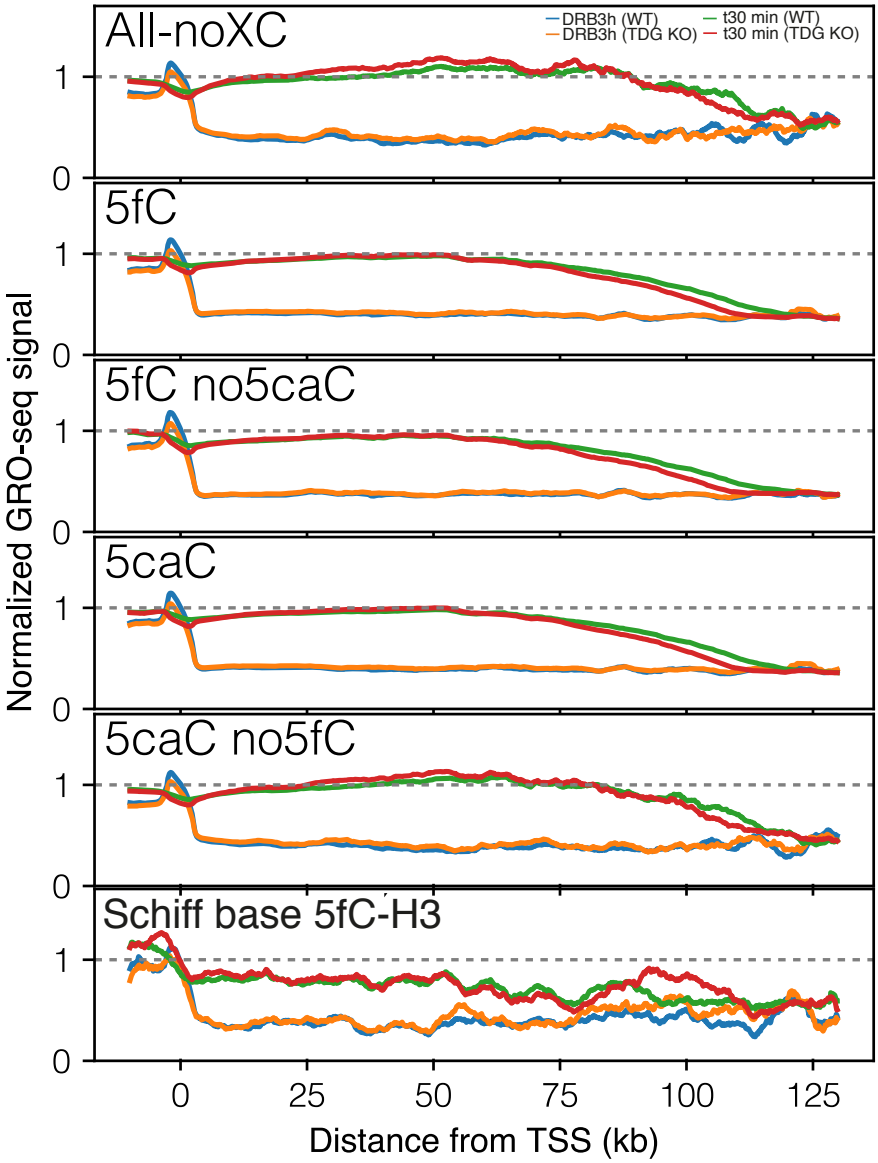
Supplementary Fig. 9

(a) Volcano plots showing log2 fold changes against log10 FDR for individual stalling sites using the forward primer (left panel) and reverse primer (right panel) (FDR was calculated by a Benjamini-Hochberg correction on exact p-value from negative binomial distribution). **(b)** To compute the closest lysine to each possible nucleotide in nucleosome, distance calculation was done via a switching function (saturation-like curve) such that anything below 0.5 nm gives a value of 1, and anything above that decays to 0 (at ~1 nm is almost zero). Experiments were repeated twice independently with similar results (n = 2).



Supplementary Fig. 10

Oligonucleotides (ODN) containing either 5fC or an abasic site and ODN after incubation with lysine and reduction using $NaCNBH_3$ (green dots) were used for qPCR analysis to compare amplification efficiency using the polymerase Q5 (see Supplementary Table 1 for sequence details). The error bars represent the standard deviation of the mean from two independent experiments (n=2).



Supplementary Fig. 11

Metagene analysis of normalized GRO-Seq signal at “all” (without 5fC/5caC) genes and genes containing 5fC, 5caC, 5fC (but not 5caC), 5caC (but not 5fC) or crosslinked 5fC/H3 sites. See supplementary methods ‘Changes in time dependent Pol II transcription elongation in the presence of 5fC’ for further details.

Supplementary Tables

	Widom 601	ODN (abasic site)
Template (5'-3')	ATCGAGAATCCCGGTGCCGA GGCCGCTCAATTGGTCGTAG ACAGCTCTAGCACCGCTTAA ACGCACGTACGCGCTGTCCC CCGCGTTTTAACCGCCAAGG GGATTACTCCCTAGTCTCCA GGCACGTGTCAGATATATACA TCCGAT	CACACCGCCAGCCACAGC AACGAACGUGCAGCGCCC CTCACGCCACAGAACATC GCATTTACGACGATTGAT GTACTAAATAGTGGGTGG TCGGTTCGCG
Primer fwd (5'-3')	ATC GAG AAT CCC GGT GCC GA	CACACCGCCAGCCACAGC AA
Primer rev (5'-3')	ATC GGA TGT ATA TAT CTG ACA CGT GCC TGG AGA	CGCGAACCGACCACCCACTA

Supplementary Table 1

DNA sequences and primers used in the study.

dxCtp concentration used for PCR

	100%	50%	20%	10%	1%
dCtp	-	100 uM	120 uM	168 uM	192.7 uM
d5fCtp	200 uM	100 uM	80 uM	32 uM	7.3 uM

Supplementary Table 2

Ratio of dNTPs for the generation of different density modified 5fC-containing DNA by PCR.

	Number of mapped reads
Hindbrain repl1 (in vitro)	44669954
Hindbrain repl2 (in vitro)	26368989
Heart repl1 (in vitro)	1665767
Heart repl2 (in vitro)	42704825
Hindbrain repl1 (in vivo)	137902491
Hindbrain repl2 (in vivo)	373859893
Heart repl1 (in vivo)	500345433
Heart repl2 (in vivo)	154280008
H3 ChIP-seq repl1	98499704
H3 ChIP-seq repl2	343900483

Supplementary Table 3

Number of mapped sequencing reads.

References

1. Iurlaro, M. *et al.* In vivo genome-wide profiling reveals a tissue-specific role for 5-formylcytosine. *Genome Biol.* **17,** 141 (2016).
2. Kunz, C. *et al.* Base Excision by Thymine DNA Glycosylase Mediates DNA-Directed Cytotoxicity of 5-Fluorouracil. *PLoS Biol.* **7,** e1000091 (2009).
3. Hess, B., Kutzner, C., van der Spoel, D. & Lindahl, E. GROMACS 4: Algorithms for Highly Efficient, Load-Balanced, and Scalable Molecular Simulation. *J. Chem. Theory Comput.* **4,** 435–447 (2008).
4. Darden, T., York, D. & Pedersen, L. Particle mesh Ewald: An N ·log(N) method for Ewald sums in large systems. *J. Chem. Phys.* **98,** 10089–10092 (1993).
5. Miyamoto, S. & Kollman, P. A. Settle: An analytical version of the SHAKE and RATTLE algorithm for rigid water models. *J. Comput. Chem.* **13,** 952–962 (1992).
6. Hess, B. P-LINCS: A Parallel Linear Constraint Solver for Molecular Simulation. *J. Chem. Theory Comput.* **4,** 116–122 (2008).
7. Feenstra, K. A., Hess, B. & Berendsen, H. J. C. Improving efficiency of large time-scale molecular dynamics simulations of hydrogen-rich systems. *J. Comput. Chem.* **20,** 786–798 (1999).
8. Bussi, G., Donadio, D. & Parrinello, M. Canonical sampling through velocity rescaling. *J. Chem. Phys.* **126,** (2007).
9. Berendsen, H. J. C., Postma, J. P. M., van Gunsteren, W. F., DiNola, A. & Haak, J. R. Molecular dynamics with coupling to an external bath. *J. Chem. Phys.* **81,** 3684–3690 (1984).
10. Best, R. B. & Hummer, G. Optimized molecular dynamics force fields applied to the helix-coil transition of polypeptides. *J. Phys. Chem. B* **113,** 9004–15 (2009).
11. Pérez, A. *et al.* Refinement of the AMBER force field for nucleic acids: improving the description of alpha/gamma conformers. *Biophys. J.* **92,** 3817–29 (2007).
12. Smith, D. E. & Dang, L. X. Computer simulations of NaCl association in

polarizable water. *J. Chem. Phys.* **100,** 3757–3766 (1994).

13. Collepardo-Guevara, R. *et al.* Chromatin unfolding by epigenetic modifications explained by dramatic impairment of internucleosome interactions: A multiscale computational study. *J. Am. Chem. Soc.* **137,** 10205–10215 (2015).
14. Davey, C. A., Sargent, D. F., Luger, K., Maeder, A. W. & Richmond, T. J. Solvent mediated interactions in the structure of the nucleosome core particle at 1.9 a resolution. *J. Mol. Biol.* **319,** 1097–113 (2002).
15. Li, H. & Durbin, R. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* **25,** 1754–1760 (2009).
16. Ramírez, F. *et al.* deepTools2: a next generation web server for deep-sequencing data analysis. *Nucleic Acids Res.* **44,** W160–W165 (2016).
17. Zhang, Y. *et al.* Model-based Analysis of ChIP-Seq (MACS). *Genome Biol.* **9,** R137 (2008).
18. Chen, W. *et al.* Improved nucleosome-positioning algorithm iNPS for accurate nucleosome positioning from sequencing data. *Nat. Commun.* **5,** 4909 (2014).
19. Chen, K. *et al.* DANPOS: Dynamic analysis of nucleosome position and occupancy by sequencing. *Genome Res.* **23,** 341–351 (2013).
20. Kim, D., Langmead, B. & Salzberg, S. L. HISAT: a fast spliced aligner with low memory requirements. *Nat. Methods* **12,** 357–360 (2015).
21. Anders, S., Pyl, P. T. & Huber, W. HTSeq--a Python framework to work with high-throughput sequencing data. *Bioinformatics* **31,** 166–169 (2015).
22. Robinson, M. D., McCarthy, D. J. & Smyth, G. K. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics* **26,** 139–140 (2010).
23. Wang, L. *et al.* Molecular basis for 5-carboxycytosine recognition by RNA polymerase II elongation complex. (2015). doi:10.1038/nature14482
24. Shen, L. *et al.* Genome-wide Analysis Reveals TET- and TDG-Dependent 5-Methylcytosine Oxidation Dynamics. *Cell* **153,** 692–706 (2013).