Molecular basis for 5-carboxycytosine recognition by RNA polymerase II elongation complex

Lanfeng Wang1, Yu Zhou2, Liang Xu1, Rui Xiao2, Xingyu Lu3, Liang Chen2, Jenny Chong3, Hairi Li2, Chuan He3, Xiang-Dong Fu* & Dong Wang1

DNA methylation at selective cytose residues (5-methylcytosine (5mC)) and their removal by TET-mediated DNA demethylation are critical for setting up pluripotent states in early embryonic development1–3. TET enzymes successively convert 5mC to 5-hydroxymethylcytosine (5hmC), 5-formylcytosine (5fC), and 5-carboxylcytosine (5caC), with 5fC and 5caC subject to removal by thymine DNA glycosylase (TDG) in conjunction with base excision repair4–6. Early reports indicate that 5fC and 5caC could be stably detected on enhancers, promoters and gene bodies, with distinct effects on gene expression, but the mechanisms have remained elusive4–6. Here we determined the X-ray crystal structure of yeast elongating RNA polymerase II (Pol II) in complex with a DNA template containing oxidized 5mCs, revealing specific hydrogen bonds between the 5-carboxyl group of 5caC and the conserved epi-DNA recognition loop in the polymerase. This causes a positional shift for incoming nucleoside 5'-triphosphate (NTP), thus compromising nucleotide addition. To test the implication of this structural insight in vivo, we determined the global effect of increased 5fC/5caC levels on transcription, finding that such DNA modifications indeed retarded Pol II elongation on gene bodies. These results demonstrate the functional impact of oxidized 5mCs on gene expression and suggest a novel role for Pol II as a specific and direct epigenetic sensor during transcription elongation.

Epigenetic DNA methylation (5mC) is an important regulator of gene transcription recognized by several families of protein readers, such as methyl-CpG-binding domain proteins (MBDs) and ubiquitin-like PHD and RING finger domain-containing proteins (for example, UHRF1), and certain zinc-finger proteins (kaiso; also known as ZBTB33)9,10. TET enzymes iteratively oxidize 5mC to 5hmC, 5fC, and 5caC3–6, and TDG coupled with base excision repair further promotes the removal reactions (Fig. 1c)14. The crystal structure (EC-I) revealed that the upstream RNA/DNA hybrid region maintains a post-translocation state register in which the active site is empty and ready for NTP loading (Fig. 1d). About 50% of 5caC nucleobase (yellow coloured in Fig. 1d, h, see also Extended Data Fig. 1a, b) accommodates at a new translocation intermediate position, located about halfway between the canonical i+1 and i+2 sites. The other 50% of 5caC nucleobase is partially inserted into the i+1 position (cyan coloured in Fig. 1d, g). Importantly, we detected specific hydrogen bonds between the 5-carboxyl moiety of 5caC and the side chain of residue Q531 at a loop in the fork region of Rpb2 (the second largest subunit) (Fig. 1e, f)6. We termed it the ‘epi-DNA recognition loop’ or ‘fork loop 3’, because it recognizes the epigenetic DNA modification in the major groove and is next to the previously identified fork loop 1 and fork loop 2 within the fork region6. The specific hydrogen-bonding interactions with 5caC result in a 90-degree rotation of the side chain of Q531 of the yeast Pol II Rpb2, switching its interacting partner in the upstream RNA/DNA hybrid region17 to the nucleobase of 5caC at i+1 position register (Fig. 1e, f). This causes 5caC to shift into a new translocation intermediate position right above the bridge helix (Fig. 1e, f, h), which we termed the ‘midway position’.

To investigate the potential impact of 5caC on nucleotide incorporation, we next solved the structure of the Pol II EC with a 5caC at the i+1 site in the presence of a non-hydrolysable GTP analogue (GMPCPP) to mimic the state of GTP binding opposite 5caC (EC-II). We found that while 5caC forms a canonical Watson–Crick base pair with GMPCPP (Fig. 2a and Extended Data Fig. 1c, d), the base pair shifts to another translocation intermediate position, −1.5 Å away from its canonical position towards the downstream main channel (Fig. 2b, d and Extended Data Fig. 2a). The interaction between the epi-DNA recognition loop and 5caC probably causes this positional shift (Fig. 2b–d), which disrupts the proper alignment between Rpb1 L1081 and the substrate, as well as the correct positioning of the 3'-RNA terminus and the substrate that is crucial for full closure of the trigger loop and effective GTP addition17,18. The nucleobase of the substrate now misaligns with Rpb1 T831 in the bridge helix (Fig. 2a), leading to a partially open conformation of the trigger loop (Extended Data Fig. 2b).

To determine further whether the specific hydrogen-bonding interaction between the Pol II epi-DNA recognition loop (Rpb2 Q531) and 5caC (Fig. 2b, c) causes a reduction in GTP addition efficiency, we purified two yeast Pol II point mutants (Rpb2 Q531H and Q531A) and measured GTP incorporation on the 5caC template in comparison with wild-type Pol II. The Pol II Q531A mutation abolishes the specific hydrogen bonds between the side chain of residue 531 and assembled on an RNA/DNA scaffold that contains a 5caC at the i+1 site (Fig. 1b) to mimic the stage when Pol II EC encounters 5caC during transcription elongation. This scaffold recapitulated the impediment of Pol II elongation in the in vitro reactions (Fig. 1c)14. The crystal structure (EC-I) revealed that the upstream RNA/DNA hybrid region maintains a post-translocation state register in which the active site is empty and ready for NTP loading (Fig. 1d). About 50% of 5caC nucleobase (yellow coloured in Fig. 1d, h, see also Extended Data Fig. 1a, b) accommodates at a new translocation intermediate position, located about halfway between the canonical i+1 and i+2 sites. The other 50% of 5caC nucleobase is partially inserted into the i+1 position (cyan coloured in Fig. 1d, g). Importantly, we detected specific hydrogen bonds between the 5-carboxyl moiety of 5caC and the side chain of residue Q531 at a loop in the fork region of Rpb2 (the second largest subunit) (Fig. 1e, f)6. We termed it the ‘epi-DNA recognition loop’ or ‘fork loop 3’, because it recognizes the epigenetic DNA modification in the major groove and is next to the previously identified fork loop 1 and fork loop 2 within the fork region6. The specific hydrogen-bonding interactions with 5caC result in a 90-degree rotation of the side chain of Q531 of the yeast Pol II Rpb2, switching its interacting partner in the upstream RNA/DNA hybrid region17 to the nucleobase of 5caC at i+1 position register (Fig. 1e, f). This causes 5caC to shift into a new translocation intermediate position right above the bridge helix (Fig. 1e, f, h), which we termed the ‘midway position’.

To understand the molecular basis underlying the Pol II EC recognition of oximiCs, we performed structural studies of the complex

5Shaggs School of Pharmacy and Pharmaceutical Sciences, The University of California, San Diego, 9500 Gilman Drive, La Jolla, California 92093, USA. 6Department of Cellular and Molecular Medicine, School of Medicine, The University of California, San Diego, 9500 Gilman Drive, La Jolla, California 92093, USA. 7Department of Chemistry, Department of Biochemistry and Molecular Biology, and Institute for Biophysical Dynamics, Howard Hughes Medical Institute, The University of Chicago, Chicago, Illinois 60637, USA.

*These authors contributed equally to this work.

©2015 Macmillan Publishers Limited. All rights reserved
Figure 1 | Pol II directly recognizes 5caC during transcription. a, Epigenetic modification cycle of cytosine. b, The RNA/DNA scaffold used in both structural and biochemical analysis. C* indicates 5caC residue. c, Impeded Pol II elongation on the 5caC-containing template relative to the unmodified C template. Time points are 0 s, 5 s, 15 s, 30 s, 1 min, 5 min, 20 min, and 1 h (left to right). d, The overall Pol II EC structure containing a site-specific 5caC (EC-I). Colour-coded are template DNA (blue), non-template DNA (green) and RNA (red). The two 5caC conformers are highlighted in yellow and cyan, respectively. Part of the template. Time points are 0, 5 s, 15 s, 30 s, 1 min, 5 min, 20 min, and 1 h (left to right). e, Impeded Pol II elongation on the 5caC-containing template relative to the unmodified C template. The critical Gln residue (Q531 in yeast Rpb2) is conserved among several fungal species containing active TET/JBP enzymes and oxi-
mCs, such as Agaricomycetes and Pucciniomycetes, and is substituted by the functionally equivalent His residue in mammals (Extended Data Fig. 5b–d). Indeed, we observed impeded Pol II elongation by human Pol II in HeLa nuclear extracts (Extended Data Fig. 5b–d). 

Figure 2 | Interaction between 5caC and epi-DNA recognition loop compromises GTP incorporation. a, The Pol II EC structure containing a matched GMPCPP opposite the 5caC site (EC-II). The colour codes are the same as Fig. 1 except for 5caC (yellow) and GMPCPP (orange). b–d, The GMPCPP:5caC base pair is shifted towards the downstream main channel from the canonical GMPCPP:C position (PDB accession 2E2J). The side chain of Rpb2 Q531 rotates 100° to interact with 5caC. e–g, Comparison of catalytic rate constants (kpol), substrate dissociation constants (Kd,app) (g) of GTP incorporation opposite the 5caC template by wild-type (WT), Q531H and Q531A Pol II, respectively. The mean values are presented and error bars are standard deviations derived from three independent experiments.
relative to the unmodified C template\textsuperscript{4}. The conservation of this critical Gln/His residue in eukaryotes coincides with the existence of 5fC/5caC modifications. In contrast, bacteria and archaea RNA polymerases carry Ala or Pro at the corresponding position in the $\beta$ loop II region (Extended Data Fig. 5a)\textsuperscript{22,24}, and consistently, we found that 5caC has no observable effect on *Escherichia coli* RNA polymerase transcription elongation \textit{in vitro} (Extended Data Fig. 7, bottom). It is interesting to note that glycosylated cytosome derivatives are also present in some phage and bacterial DNA genomes, pointing to future investigations to understand how these modifications (bulkier than 5fC/5caC) may be recognized during transcription.

Our structural studies also shed light on the canonical Pol II translocation process by revealing two new translocation intermediate positions of the 5caC template before and after GTP binding. The first translocation intermediate position of 5caC sits above the bridge helix in the absence of GTP. Upon GTP binding, the 5caC template is shifted to a new translocation intermediate position allowing the formation of a base pair with incoming GTP. The translocation intermediate states are similar to the translocation intermediate states on an unmodified DNA template recently suggested by molecular simulation\textsuperscript{22}. Our ability to capture the crystal structures of these Pol II translocation intermediates suggests that the specific interactions between the Pol II recognition loop (Q531) and the 5-carboxyl group of 5caC stabilize the translocation intermediates that are otherwise too transient to be captured on unmodified DNA template.

Aligning the structures of 5caC-paused Pol II EC with bulky DNA lesion-arrested or $\alpha$-amanitin-arrested Pol II EC\textsuperscript{18,23,24} reveals additional insights into Pol II pausing and arrest. Notably, 5caC, CPD, pyrroplatin-dG and the i+1 transition template base in $\alpha$-amanitin-arrested Pol II EC are all accommodated above the bridge helix (Fig. 3a–c), even though their exact locations, orientations and interactions with Pol II greatly differ (see Methods). A similar (but not identical) ‘above-the-bridge-helix’ translocation intermediate has also been recently observed in an elemental paused *E. coli* RNA polymerase structure (ePEC) with a kinked bridge helix to occlude the canonical i+1 template position\textsuperscript{25}. Taken together, we propose that these observations point to a common translocation checkpoint that serves as a rate-limiting step for the transition of the DNA template nucleobase to cross over the bridge helix and subsequently insert into the canonical i+1 site to guide RNA synthesis. While DNA lesions have been proposed to interfere with Pol II elongation via steric hindrance\textsuperscript{26,27}, our current data suggest that Pol II can also directly sense epigenetically modified DNA (5caC/5fC) through specific hydrogen-bonding interactions.

We further noticed a remarkable mechanistic similarity in 5caC recognition by several unrelated family proteins. For example, residue Q369 in human Wilms tumor protein 1 (WT1) (PDB accession 4R2R)\textsuperscript{13} and residue N157 in human TDG (PDB accession 3UO7)\textsuperscript{28} are both functionally equivalent to Q/H531 residues in Pol II in recognizing the 5caC carboxyl group via specific hydrogen bonds (Fig. 3d–f). We thus speculate that 5caC could be a potential epigenetic mark for recognition by a variety of protein readers (including Pol II itself) via specific hydrogen-bonding interactions with its 5-carboxyl moiety.

To determine further the functional consequences of oxi-mCs on Pol II transcription elongation in mammalian cells, we measured the \textit{in vivo} transcription elongation rate on a pair of isogenic mouse embryonic stem (ES) cells (Tdgy\textsuperscript{-}amanitin and Tdgy\textsuperscript{-}knockout mouse ES cells derived from conditional TDG-knockout mice) by global nuclear run-on coupled with deep sequencing (GRO-seq) (Fig. 4a). Previous studies showed that, relative to wild type, TDG knockout led to a substantial increase of global 5fC/5caC levels\textsuperscript{26}. The GRO-seq experiments allowed us to measure the front edge of waves of nascent transcripts at different time points to deduce the rate of Pol II transcription elongation.

We observed retarded Pol II elongation in TDG-knockout ES cells relative to wild-type ES cells after the functional impact was sufficiently accumulated, as exemplified on the long \textit{MyoI} gene (Fig. 4b). Further metagene analysis of the middle points of wild-type and TDG-knockout mouse ES cells at different time points revealed a clear reduction of Pol II elongation in TDG-knockout relative to wild-type ES cells after 30 min of synchronized transcription, although the differences at earlier times were not evident (Fig. 4c). We next analysed the GRO-seq read density profile of 5fC/5caC read density by gene $\pm$10 kb around individual middle points followed by linear regression to determine the slope (Fig. 4d). We observed progressive slowing down of Pol II in TDG-knockout ES cells relative to wild-type cells, as indicated by decreasing slopes, and the read density ratio (TDG knock/wild type) at 30 min was significantly smaller relative to control ($P$ value $= 1.52 \times 10^{-10}$ from one-sided Kolmogorov–Smirnov test) (Fig. 4d). Finally, to determine the dosage-dependent effect of 5fC/5caC on Pol II transcription elongation, we focused on the data at 30 min and segregated genes into two groups according to increased levels of 5fC/5caC in response to TDG knockout and compared the middle points at individual assay points. The data indicate a correlation between increased 5fC/5caC and a reduced transcription elongation rate among genes in group 2 (high 5fC/5caC level) relative to group 1 (low 5fC/5caC level) (Fig. 4e).

Together, these global data demonstrate retarded Pol II elongation by enhanced 5fC/5caC levels in the gene body. The combination of \textit{in vitro} and \textit{in vivo} data strongly indicates a direct impact of 5fC/5caC on Pol II elongation on the DNA template.

We present structural and biochemical evidence to suggest that Pol II has the ability to sense the DNA oxidative methylation state directly through its conserved epi-DNA recognition loop, and that it transiently slows down at oxi-mC (5fC/5caC) sites during transcription. Since 5fC/5caC are not distributed evenly across the genome and show considerable variation between cell types, it is conceivable that these pausing effects may add a new layer of fine-tuned regulation of Pol II transcription elongation dynamics. For example, compared with transcription of short genes, the cumulative consequences of pausing effects at 5fC/5caC sites probably have much more profound regulatory impacts on transcribing some long genes that are preferentially expressed in the brain and have crucial roles in neuronal integrity\textsuperscript{29}.
The transient Pol II pausing at 5fC/5caC sites may also provide signals for the recruitment of various transcription elongation factors, chromatin remodelling complexes, messenger RNA processing machineries, and TDG and the base excision repair machineries to the oxi-mC sites to induce additional functional consequences. On the basis of the similarity between direct Pol II recognition of 5caC and the role of Pol II in sensing bulky DNA lesions in transcription-coupled nucleotide excision repair, we propose that Pol II may act as a direct sensor for a variety of DNA modification and damage events to instruct distinct downstream pathways.

Online Content Methods, along with any additional Extended Data display items and Source Data, are available in the online version of the paper; references unique to these sections appear only in the online paper.

Received 19 June 2014; accepted 20 April 2015.
Published online 29 June 2015.

1. Pastor, W. A., Aravind, L. & Rao, A. Tetonic shift: biological roles of Tet proteins in DNA demethylation and transcription. Nature Rev. Mol. Cell Biol. 14, 341–356 (2013).

2. Wu, H. & Zhang, Y. Reversing DNA methylation: mechanisms, genomics, and biological functions. Cell 156, 45–68 (2014).

3. Tahiliani, M. et al. Conversion of 5-methylcytosine to 5-hydroxymethylcytosine in mammalian DNA by MLL partner TET1. Science 324, 930–933 (2009).

4. Pfaffneder, T. et al. The discovery of 5-formylcytosine in embryonic stem cell DNA. Angew. Chem. Int. Ed. Engl. 50, 7008–7012 (2011).

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

6. He, Y. F. et al. Tet-mediated formation of 5-carboxycytosine and its excision by TDG in mammalian DNA. Science 333, 1303–1307 (2011).

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

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

9. Klose, R. J. & Bird, A. P. Genomic DNA methylation: the mark and its mediators. Trends Biochem. Sci. 31, 89–97 (2006).

10. Moore, L. D., Le, T. & Fan, G. DNA methylation and its basic function. Neuropsychopharmacology 38, 23–38 (2013).

11. Spruijt, C. G. et al. Dynamic readers for 5-(hydroxymethyl)cytosine and its oxidized derivatives. Cell 152, 1146–1159 (2013).

12. Iurlaro, M. et al. A screen for hydroxymethylcytosine and formylcytosine binding proteins suggests functions in transcription and chromatin regulation. Genome Biol. 14, R119 (2013).
Hashimoto, H. et al. Wilms tumor protein recognizes 5-carboxylycitosine within a specific DNA sequence. Genes Dev. 28, 2304–2313 (2014).

Keller, M. W. et al. 5-formylcytosine and 5-carboxylycitosine reduce the rate and substrate specificity of RNA polymerase II transcription. Nature Struct. Mol. Biol. 19, 831–833 (2012).

Huang, Y. & Rao, A. New functions for DNA modifications by TET-JBP. Nature Struct. Mol. Biol. 19, 1061–1064 (2012).

Cramer, P., Bushnell, D. A. & Kornberg, R. D. Structural basis of transcription: RNA polymerase II at 2.8 Ångstrom resolution. Science 292, 1863–1876 (2001).

Walmacq, C. et al. Mechanism of translesion transcription by RNA polymerase II and its role in cellular resistance to DNA damage. Mol. Cell 46, 18–29 (2012).

Weixlbaumer, A., Leon, K., Landick, R. & Darst, S. A. Structural basis of transcriptional pausing in bacteria. Cell 152, 431–441 (2013).

Lindsey-Boltz, L. A. & Sancar, A. RNA polymerase: the most specific damage recognition protein in cellular responses to DNA damage? Proc. Natl Acad. Sci. USA 104, 13213–13214 (2007).

Hanawalt, P. C. & Spivak, G. Transcription-coupled DNA repair: two decades of progress and surprises. Nature Rev. Mol. Cell Biol. 9, 958–970 (2008).

Zhang, L. et al. Thymin DNA glycosylase specifically recognizes 5-carboxylycitosine-modified DNA. Nature Chem. Biol. 8, 328–330 (2012).

Polymenidou, M. et al. Long pre-mRNA depletion and RNA missplicing contribute to neuronal vulnerability from loss of TDP-43. Nature Neurosci. 14, 459–468 (2011).

Sarker, A. H. et al. Recognition of RNA polymerase II and transcription bubbles by XPO, CSB, and TFIH: insights for transcription-coupled repair and Cockayne syndrome. Mol. Cell 20, 187–198 (2005).

Acknowledgements D.W. acknowledges the National Institutes of Health (NIH) (GM102362), a Kimmel Scholars award from the Sidney Kimmel Foundation for Cancer Research, and start-up funds from the Skaggs School of Pharmacy and Pharmaceutical Sciences, University of California, San Diego. This work was also supported by NIH grant HG006627 and the Howard Hughes Medical Institute to C.H., and NIH grants GM052872 and HG004659 to X.-D. F. We are grateful to C. Kaplan for providing Saccharomyces cerevisiae Pol II Rpb2 Q531H and Q531A mutant strains.

Author Contributions D.W. conceived the original idea and, together with X.-D.F., designed the experiments. X.L. carried out synthesis of DNA templates. J.C., L.W. and D.W. performed crystallization, data collection and structural refinement. L.X. performed the in vitro transcription assay. Y.Z., R.X., C.H. and D.W. performed the in vivo GRO-seq assay. L.W., Y.Z., L.X., X.L., J.C., C.H., X.-D.F. and D.W. wrote the paper.

Author Information GRO-seq data have been deposited in the Gene Expression Omnibus database under accession GSE54748. Atomic coordinates and structure factors for the reported crystal structures have been deposited in the Protein Data Bank under accessions 4Y52 and 4Y7N for EC-I and EC-II, respectively. Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Readers are welcome to comment on the online version of the paper. Correspondence and requests for materials should be addressed to D.W. (dongwang@ucsd.edu) or X.-D.F. (xdfu@ucsd.edu).
Nonlinear-regression data fitting was performed using Prism 6. The time dependence of product formation was fit to a one-phase association equation (product = A(1−e−kt+C)+C) to determine the observed rate (k detached). The substrate concentration dependence was fit to a hyperbolic equation (k detached = k max [substrate] / (K M + [substrate])) to obtain values for the maximum rate of NTP incorporation (k max) and apparent Km (K M) governing NTP binding essentially as described. The specificity constant was determined by k detached/KM app.

Cell culture and in vivo transcription rate measurement. Wild-type mouse ES cells (Tgδ4−/− and TgδGκ−/−) and knockout mouse ES cells (Tgδ4−/−) were cultured in KnockoutDMEM (Life Technologies, catalogue no. 10828-018) supplemented with 15% KnockoutDMEM Serum Replacement (Life Technologies, catalogue no. 10828-028), 2 mM l-glutamine (Life Technologies, catalogue no. 2530-081), 1× non-essential amino acids (Life Technologies, catalogue no. 11140-050), 1× penicillin-streptomycin (Life Technologies, catalogue no. 15140-122), 0.1 mM 2-mercaptoethanol (Life Technologies, catalogue no. 21985-023), 1,000 U ml−1 LIF (Millipore, catalogue no. ESG1106), 3 μM CHIR99021 (Stemgent, catalogue no. 04-0004) and 1 μM PD0325901 (Stemgent, catalogue no. 04-0006). The DRB releasing GRO-seq assays were carried out in mouse ES cells under both wild-type and TgδGκ-knockout conditions. For each time-course assay, there are five samples prepared for GRO-seq: (1) NODRB (without DRB treatment); (2) DRB3H (DRB treatment for 3 h, and this is the 0 time point); (3) 10M (10 min after washing out DRB); (4) 20M (20 min after washing); (5) 30M (30 min after washing). For DRB treatment, we grew cells in a 10 cm plate to 70–80% confluence, treated cells by addition of DRB (Sigma) at a final concentration of 100 μM to the culture medium and incubated for 3 h in the incubator, removed DRB by quick washing cells three times with PBS, then incubated in fresh medium in the incubator to different time points. GRO-seq was implemented as previously described35,36, and the GRO-seq libraries were subjected to medium in the incubator to different time points. GRO-seq was implemented as previously described35,36, and the GRO-seq libraries were subjected to

RESEARCH LETTER

Preparation of Pol II ECs. Saccharomyces cerevisiae Pol II was purified as previously described37. PAGE-purified RNA oligonucleotides were purchased from Thermo Fisher Scientific, non-template DNA oligonucleotides were obtained from IDT, and template DNA oligonucleotides with 5caC were prepared and purified as previously described4. The template DNA, non-template DNA and RNA oligonucleotides were annealed to form the scaffold. To form the Pol II EC, Pol II was mixed with scaffold in the reaction buffer (20 mM Tris (pH 7.5), 40 mM KCl and 5 mM dithiothreitol (DTT)). The final concentrations were 2 μM Pol II, 10 μM template DNA and 20 nM non-template DNA and RNA oligonucleotides. The mixture was incubated at room temperature for 1 h, followed by ultrafiltration to remove excess oligonucleotides. The Pol II elongation complex was crystallized using the hanging drop method and from solutions containing 390 mM (NH4)2HPO4, 10 mM Na2HPO4, pH 6.0, 50 mM dioxane, 10 mM DTT and 10.7–11.6% (w/v) PEG6000. Crystals were transferred in a stepwise manner to cryo buffer as previously described47. For the Pol II EC with GMPCPP, Pol II EC crystals were soaked with 5–10 mM GMPCPP and 10 mM MgCl2 overnight before harvest.

Data collection and structure determination of 5caC-paused Pol II ECs. Diffraction data were collected on beamlines 8.2.1 and 50.2 at the Advanced Light Source, Lawrence Berkeley National Laboratory. Data were processed in DENOVO and SCALPACK (HR2300)58. Model building was performed with the program Coot 32, and refinement was done with REFMAC5 with TLS (CCP4i) or PHENIX (Extended Data Table 1)39. Electron density maps are shown in Extended Data Fig. 1. EC-II refers to the Pol II EC crystal structure that contains a site-specific 5caC at the i+1 site in the absence of GTP binding. EC-II refers to the Pol II EC crystal structure that contains a site-specific 5caC at the i+1 site in the presence of GMPPCP. Ramachandran plot of EC-I showed 85.57%, 11.70% and 2.73% of EC-I residues are in allowed, allowed and disallowed regions, respectively. For EC-II, 86.00%, 11.22%, and 2.78% of residues are in above regions, respectively. All structural models in the figures were superimposed with the bridge helix region (Rpb1 822–840) near the active site using Coot 42 and PyMOL43.

Pol II purification for in vitro transcription assays. S. cerevisiae Pol II and mutants were purified essentially as previously described42. Briefly, Pol II (with recombinant protein A tag at Rpb3 subunit) was first affinity-purified by IgG column. The Pol II elution from IgG column was further purified using HiTrap Heparin and Mono Q (GE Healthcare). The final pure Pol II (Extended Data Fig. 3d) was ready for future in vitro transcription experiments.

In vitro transcription assays. The S. cerevisiae Pol II ECs for transcription assays were assembled using established methods46. Briefly, an aliquot of 5′-32P-labeled RNA was annealed with a 1.5-fold amount of template DNA and a 2-fold amount of non-template DNA to form RNA/DNA scaffold in elongation buffer (20 mM Tris-HCl, pH 7.5, 40 mM KCl and 5 mM MgCl2). An aliquot of annealed scaffold of RNA/DNA was then incubated with a fourfold excess amount of Pol II at room temperature for 10 min to ensure the formation of Pol II EC. The in vitro transcription started when the Pol II EC was mixed with equal volumes of GTP solution. The final concentrations were 25 nM scaffold, 100 nM Pol II and 1 μM GTP in the elongation buffer. Reactions were quenched at various time points by the addition of one volume of 0.5 M EDTA (pH 8.0). (Time points are 0, 5, 15, 30 s, 1 min, 5 min, 20 min, and 1 h). The quenched products were analysed by denaturing PAGE and visualized using a storage phosphor screen and Phosphor FX imager (Bio-Rad). The in vitro transcription assay of E. coli RNA polymerase (RNAP, New England Biolabs (NEB)) was performed using the same procedure as S. cerevisiae RNA Pol II transcription.

For the transcription of human Pol II in the nuclear extract of HeLa cells (Life Technologies), the excess annealed scaffold was incubated with nuclear extract of HeLa cells for 5 min before the addition of α-32P-GTP. The final concentrations were 1 μM scaffold, 1 μM α-32P-GTP (0.2 μCi μl−1) and 3 mg ml−1 protein of nuclear extract. Reactions were then quenched at various time points by the addition of one volume of 0.5 M EDTA (pH 8.0). The quenched products were analysed by denaturing PAGE and visualized using a storage phosphor screen and Phosphor FX imager (Bio-Rad). All transcription assays described earlier were performed independently in triplicates.

In vitro RNA Pol II transcription kinetic assay and analysis. The assay was carried out as previously described44. Briefly, nucleotide incorporation assays were conducted by pre-incubating 50 nM annealed scaffold containing a site-specific 5caC modification at the template with 200 nM purified Pol II (wild type, Q531H and Q531A) for 10 min in elongation buffer at room temperature. The Pol II EC was then mixed with an equal volume of solution containing 40 mM KCl, 20 mM Tris-HCl (pH 7.5), 10 mM DTT, 10 mM MgCl2, 0.1 mM each of α-32P-GTP, and various nucleotides. Final reaction concentrations after mixing were 25 nM scaffold, 100 nM Pol II, 5 mM MgCl2 and various nucleotide concentrations in elongation buffer. Reactions were quenched at various times by addition of one volume of 0.5 M EDTA (pH 8.0) and analysed by denatured PAGE.

To measure the concordance between replicated samples, we counted the number of the GRO-seq reads in all annotated genes (UCSC refGene) and did pairwise comparisons38. Transcripts with the same start and end positions were used once. Having established the data reproducibility, we combined replicated data sets for comparison between biological conditions at different assay points.

To estimate the Pol II elongation rates, we calculated the metagene profiles for all assay points. Only the genes with RPMK ≥ 0.5 in the NODRB sample were kept for meta-analysis. The genes were aligned at TSSs, and mapped reads were counted in 100 bp bins across the gene bodies. The counts were normalized to one million total reads per sample, and were averaged for each bin by the number of covering genes and normalized by the relative gene expression in the NODRB sample. The meta-profiles from the normalized counts were smoothed with a 1 kb moving window. The middle point of the ensemble transcription wave at each time point after washing DRB was computed as the position at which the signal reached half of that in the NODRB control.

To compare elongation differences at different assay points on individual genes, we calculated the GRO-seq read density in ±10 kb window around the middle points identified in wild-type mouse ES cell lines. At each assay point (10M, 20M, 30M), the counts for each gene in wild-type and TgδGκ-knockout conditions were pairwise compared and linear regressions were fitted to check the trend of change. The samples without DRB treatment were used as control. The changes of 5caC/5caC levels on genes were calculated as the differences of normalized ChiP-seq signals under 5caC/Sac5C ChiP-seq peaks from knockout to wild type based on the published ChiP-seq data from ref. 8, and the genes were divided into two groups with average 5caC level of high increased and 5caC level of low increased.

Comparison of 5caC-paused, DNA lesion-arrested, and ω-aminatinted arrested Pol II EC. All the structures were aligned by superimposition of the Pol II bridge helix region (residues 822–840 in Rpb1). The i+1 5caC, CPD, pyrRiolipin-DG and i+1 transition template base in ω-aminatinated-arrested Pol II EC is all accommodated above the bridge helix, even though their exact locations, orientations and interactions with Pol II generally vary. ω-aminatinated appears to capture the Pol II transcription intermediate indirectly by jamming the movement of the Pol II bridge helix and trapping the trigger loop in an inactive conformation, whereas the conformation of CPD and pyrRiolipin-DNA lesions is largely governed by their coordinate tug of war with the bridge helix. In contrast to all of these previous cases, the translocation intermediate of 5caC nucleobase forms a direct interaction with the Pol II epi-DNA recognition loop. Second, the upstream RNA/DNA hybrid adopts a substantially tilted for CPD-lesion-
arrested Pol II EC. Finally, while 5caC and pyriplatin-DNA lesion can form Watson–Crick base pairs with incoming NTP, thus allowing template-dependent nucleotide addition, the CPD-DNA lesion fails to form such a base pair with the incoming nucleotide, therefore only allowing template-independent ATP incorporation. In contrast to all of these previous cases, the translocation intermediate of the 5caC nucleobase forms a direct interaction with the Pol II epi-DNA recognition loop.

31. Otwinowski, Z. & Minor, W. Processing of X-ray diffraction data collected in oscillation mode. Methods Enzymol. **276**, 307–326 (1997).
32. Emsley, P. & Cowtan, K. Coot: model-building tools for molecular graphics. Acta Crystallogr. D **60**, 2126–2132 (2004).
33. Adams, P. D. et al. PHENIX: a comprehensive Python-based system for macromolecular structure solution. Acta Crystallogr. D **66**, 213–221 (2010).
34. DeLano, W. L. The PyMOL Molecular Graphics System (DeLano Scientific, 2002).
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).
36. Wang, D. et al. Reprogramming transcription by distinct classes of enhancers functionally defined by eRNA. Nature **474**, 390–394 (2011).
37. Langmead, B., Trapnell, C., Pop, M. & Salzberg, S. L. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. Genome Biol. **10**, R25 (2009).
38. Karolchik, D. et al. The UCSC Genome Browser database: 2014 update. Nucleic Acids Res. **42**, D764–D770 (2014).
Extended Data Figure 1 | Electron density maps of Pol II EC-I and EC-II.  

(a) $2F_o - F_c$ map (blue) of Rpb2 Q531 in epi-DNA recognition loop and the opposite 5caC in Pol II EC-I, contoured at 1.0σ.  

(b) $F_o - F_c$ omit map (green) of Pol II EC-I (with 5caC omission), contoured at 3.0σ.  

(c) $2F_o - F_c$ map (blue) of GMPCPP paired with 5caC in Pol II EC-II, contoured at 1.0σ.  

(d) $F_o - F_c$ omit map (green) of Pol II EC-II (with GMPCPP and 5caC omission), contoured at 3.0σ.
Extended Data Figure 2 | Structural comparison between Pol II EC-I, EC-II and Pol II EC containing unmodified C template and a matched GTP.

a, Superimposition of Pol II EC-I and EC-II structures. Rpb2 Q531 and 5caC in EC-II are in magenta to differentiate between those counterparts in EC-I. These two structures are aligned using the bridge helix (BH) region (Rpb1 822–840).

b, Superposition of Pol II EC-II containing 5caC template and GMPCPP with Pol II EC with closed trigger loop (TL; containing unmodified C template and GTP; PDB accession 2E2H). The two structures are aligned using the bridge helix region (Rpb1 822–840).
Extended Data Figure 3 | Kinetic study of GTP incorporation opposite 5caC template by purified Pol II proteins. a–c, Representative kinetic parameter fitting curves from three independent experiments for GTP incorporation opposite 5caC template for Pol II wild type (WT; a), Pol II Q531H (b) and Pol II Q531A (c). d, Purified Pol II wild-type, Pol II Q531H and Pol II Q531A proteins used in the *in vitro* transcription experiments.
Extended Data Figure 4 | Modelling potentially similar interactions for recognition of 5fC and 5caC templates, but not for 5hmC, 5mC and C templates. a, Hydrogen bonds (black dotted lines) between Rpb2 Q531, 5caC and GMPCPP in EC-II. b, Model of the interaction between Pol II EC with 5fC template through the same hydrogen-bond interaction network. c, Model of Pol II EC with 5hmC template reveals no obvious hydrogen bonding between Q531 and 5hmC. The 5hmC nucleotide structure was based on PDB accession 4R2C. d, Model of Pol II EC with 5mC template. e, Model of Pol II EC with unmodified C template. The above models were derived from the Pol II EC-II structure.
Extended Data Figure 5 | Sequence alignment of Pol II epi-DNA recognition loop across different species. a, Pol II epi-DNA recognition loop (Rpb2 521–541) is conserved from fungi to human and strictly conserved among several fungal species, highlighted with magenta dotted rectangle, which contain active TET/JBP enzymes. Key residues in the loop are highlighted in the green box. b, Hydrogen bonds (black dotted lines) between yeast Pol II Rpb2 Q531, 5caC and GMPCPP in EC-II. c, Model of human Pol II with the functionally equivalent His substitution based on EC-II structure. d, Comparison between Q531 and H531 substitution reveals the similar hydrogen-bonding interaction.
Extended Data Figure 6 | Human Pol II slows down at 5caC template in comparison with unmodified template in the context of HeLa nuclear extract. The relative transcription elongation rate is normalized by the transcription elongation rate ($k_{obs}$) from unmodified template. The relative rates from unmodified template and 5caC template are coloured in black and grey, respectively. The error bars are standard deviations derived from three independent experiments.
Extended Data Figure 7 | Comparison transcription on 5caC template with unmodified template using purified yeast Pol II and E. coli RNAP. Top, comparison of yeast Pol II; bottom, comparison of E. coli RNAP. Time points are 0, 5 s, 15 s, 30 s, 1 min, 5 min, 20 min, and 1 h (left to right). The top panel is identical to Fig. 1c and is placed here for direct comparison. nt, nucleotides.
Extended Data Figure 8 | Correlation between two replicates of GRO-seq data sets at different assay points. GRO-seq replicates (−1 and −2) were pairwise compared gene by gene on the normalized number of reads for wild-type (WT; left) and TDG-knockout (KO; right) samples. The colours show the density of points or genes. The Pearson correlation coefficients were calculated from the points and are shown on the top of each subfigure. rpm, reads per million total reads.
| Data collection                  | EC-I                      | EC-II                      |
|--------------------------------|---------------------------|----------------------------|
| Space group                    | C2                        | C2                         |
| Cell dimensions                |                           |                            |
| a, b, c (Å)                    | 166.7, 221.6, 192.4       | 168.2, 222.6, 192.8        |
| α, β, γ (°)                    | 90, 100.4, 90             | 90, 101.6, 90              |
| Resolution (Å)                 | 50.3.5 (3.56-3.5)         | 50.3.3 (3.36-3.3)          |
| R<sub>free</sub>               | 0.143 (0.583)             | 0.153 (0.762)              |
| d<sub>all</sub>                | 8.1 (1.7)                 | 9.2 (1.1)                  |
| Completeness (%)               | 94.3 (72.8)               | 99.4 (96.5)                |
| Redundancy                     | 3.6 (3.3)                 | 3.7 (3.3)                  |

| Refinement                     |                           |                            |
| Resolution (Å)                 | 49.3-3.5                  | 48.9-3.3                   |
| No. reflections               | 81,638                    | 105636                     |
| R<sub>free</sub> / R<sub>free</sub> | 20.1/23.2                | 20.7/25                    |
| No. atoms                     |                           |                            |
| Protein/Nucleic acid          | 29180                     | 29151                      |
| Ligand/ions                   | 9                         | 42                         |
| Water                         |                           |                            |
| B-factors                     |                           |                            |
| Protein/Nucleic acid          | 94.3                      | 102.8                      |
| Ligand/ions                   | 135.8                     | 95.4                       |
| Water                         |                           |                            |
| R.m.s deviations              |                           |                            |
| Bond lengths (Å)              | 0.009                     | 0.009                      |
| Bond angles (°)               | 1.355                     | 1.326                      |

*Values in parentheses are for the highest-resolution shell.