Epigenetic reprogramming by TET enzymes impacts co-transcriptional R-loops

João C. Sabino¹, Madalena R. de Almeida¹, Patrícia L. Abreu¹, Ana M. Ferreira¹, Paulo Caldas²,³, Marco M. Domingues¹, Nuno C. Santos¹, Claus M. Azzalin¹, Ana R. Grosso²,³, Sérgio F. de Almeida¹*

Affiliations: ¹Instituto de Medicina Molecular João Lobo Antunes, Faculdade de Medicina da Universidade de Lisboa, Lisboa, Portugal.
²Associate laboratory i4HB – Institute for Health and Bioeconomy, NOVA School of Science and Technology, Universidade Nova de Lisboa, 2829-516 Caparica, Portugal.
³UCIBIO-REQUIMTE, Applied Molecular Biosciences Unit, Department of Life Sciences, NOVA School of Science and Technology, Universidade Nova de Lisboa, 2829-516 Caparica.
*Correspondence to: sergioalmeida@medicina.ulisboa.pt

Abstract

DNA oxidation by ten-eleven translocation (TET) family enzymes is essential for epigenetic reprogramming. The conversion of 5-methylcytosine (5mC) into 5-hydroxymethylcytosine (5hmC) initiates developmental and cell-type-specific transcriptional programs through mechanisms that include changes in the chromatin structure. Here, we show that the presence of 5hmC in the transcribed gene promotes the annealing of the nascent RNA to the template DNA strand, leading to the formation of an R-loop. Depletion of TET enzymes reduced global R-loops in absence of gene expression changes, whereas CRISPR-mediated tethering of TET to an active gene promoted the formation of R-loops. The genome-wide distribution of 5hmC and R-loops show a positive correlation in mouse and human stem cells and overlap in half of all active genes. Moreover, R-loop resolution leads to differential expression of a subset of genes that are involved in crucial events during stem cell proliferation. Altogether, our data reveal that epigenetic reprogramming via TET activity promotes co-transcriptional R-loop formation, disclosing new mechanisms of gene expression regulation.
INTRODUCTION

During transcription, the nascent RNA molecule can hybridize with the template DNA and form a DNA:RNA hybrid and a displaced DNA strand. These triple-stranded structures, called R-loops, are physiologically relevant intermediates of several processes, such as immunoglobulin class-switch recombination and gene expression. However, non-scheduled or persistent R-loops constitute an important source of DNA damage, namely DNA double-strand breaks (DSBs). To preserve genome integrity, cells possess diverse mechanisms to prevent the formation of R-loops or resolve them. R-loop formation is restricted by RNA-binding proteins and topoisomerase 1, whereas R-loops are removed by ribonucleases and helicases (reviewed in). The ribonuclease H enzymes RNase H1 and RNase H2 degrade R-loops by digesting the RNA strand of the DNA:RNA hybrid. DNA and RNA helicases unwind the hybrid and restore the double-stranded DNA (dsDNA) structure. Several helicases unwind R-loops at different stages of the transcription cycle and in distinct physiological contexts. For instance, we previously reported that the DEAD-box helicase 23 (DDX23) resolves R-loops formed during transcription elongation to regulate gene expression programs and prevent transcription-dependent DNA damage. Intrinsic features of the transcribed DNA also influence its propensity to form R-loops. The presence of introns, for instance, prevents unscheduled R-loop formation at active genes. An asymmetrical distribution of guanines (G) and cytosines (C) nucleotides in the DNA duplex also influences R-loop propensity, with an excess of Cs in the template DNA strand (positive G:C skew) favouring R-loop formation. Moreover, chromatin and DNA features such as histone modifications, DNA-supercoiling and G-quadruplex structures also affect R-loop establishment. R-loops can also drive chromatin modifications. Promoter-proximal R-loops enhance the recruitment of the Tip60–p400 histone acetyltransferase complex and inhibit the binding of polycomb repressive complex 2 and histone H3 lysine-27 methylation. R-loops formed over G-rich terminator elements promote histone H3 lysine-9 dimethylation, a repressive mark that reinforces RNA polymerase II pausing during transcription termination.

Besides affecting histone modifications, R-loops also act as barriers against DNA methylation spreading into active genes. DNA methylation, namely 5-methylcytosine (5mC), results from the covalent addition of a methyl group to the carbon 5 of a C attached to a G through a phosphodiester bond (CpG). The activity of DNA methyltransferase (DNMT) enzymes makes 5mC widespread across the mammalian genome where it plays major roles in imprinting, suppression of retrotransposon silencing and gene expression. More than 70% of
all human gene promoters contain stretches of CpG dinucleotides, termed CpG islands (CGIs), whose transcriptional activity is repressed by CpG methylation\textsuperscript{11,12}. R-loops positioned near promoters of active genes maintain CGIs in an unmethylated state\textsuperscript{9}, likely by reducing the affinity of DNMT1 binding to DNA\textsuperscript{13}, or recruiting ten-eleven translocation (TET) methylcytosine dioxygenases\textsuperscript{14}.

The TET enzymes family members share the ability to oxidize 5mC to 5-hydroxymethylcytosine (5hmC)\textsuperscript{15-16}. 5hmC is a relatively rare DNA modification found across the genome much less frequently than 5mC\textsuperscript{17}. Genome-wide, 5hmC is more abundant at regulatory regions near transcription start sites (TSSs), promoters and exons, consistent with its role in gene expression regulation\textsuperscript{18}. The levels of 5hmC are enriched at active promoter regions, as observed upon activation of neuronal function-related genes in neural progenitors and neurons\textsuperscript{15,19}. 5hmC has the potential to modify the DNA helix structure by favouring DNA-end breathing motion, a dynamic feature of the protein–DNA complexes thought to control DNA accessibility\textsuperscript{17}. Moreover, 5hmC weakens the interaction between DNA and nucleosomal H2A-H2B dimers, facilitating RNA polymerase II elongation, and diminishes the thermodynamic stability of the DNA duplex\textsuperscript{17}. While 5mC increases the melting temperature, 5hmC reduces the amount of energy needed to separate the two strands of the DNA duplex\textsuperscript{20,21}. Molecular dynamics simulations revealed that the highest amplitude of GC DNA base-pair fluctuations is observed in the presence of 5hmC, whereas 5mC yielded GC base pairs (bp) with the lower amplitude values\textsuperscript{21}. The presence of 5hmC destabilizes GC pairing by alleviating steric constraints through an increase in molecular polarity\textsuperscript{21}.

Because features that destabilize the DNA duplex, such as supercoiling or G-quadruplexes, are known to facilitate nascent RNA annealing with the template DNA strand, we reasoned that 5hmC may favour R-loop formation. Here, we show that 5hmC promotes R-loop formation during \textit{in vitro} transcription of DNA templates. \textit{In vivo}, depletion of TET enzymes reduces R-loop levels, whereas targeting the enzyme to an active gene drives R-loop formation. Analysis of genome-wide distribution profiles shows a positive correlation between 5hmC and R-loops in mouse embryonic stem (mES) and in human embryonic kidney 293 (HEK293) cells, with a clear overlap of 5hmC and R-loops in approximately half of all active genes. We also show that 5hmC-rich regions are characterized by increased levels of phosphorylated histone H2AX ($\gamma$H2AX), a marker of DNA damage. Finally, by determining the
pathways more significantly affected by R-loops formed at 5hmC loci, we disclose novel links between R-loops and gene expression programs of stem cells.

RESULTS

Transcription through 5hmC-rich DNA favours R-loop formation

To assess the impact of cytosine methylation on R-loop formation, we performed in vitro T7 transcription of DNA fragments containing either native or modified cytosine deoxyribonucleotides (dCTPs). We synthesized three distinct DNA transcription templates, each composed of a T7 promoter followed by a 400bp sequence containing a genomic region prone to form R-loops in vivo. Two of these sequences (ACTB P1 and ACTB P2) are from the transcription termination region of the human β-actin coding gene (ACTB); the third sequence is from the human APOE gene. The DNA templates for the in vitro transcription reactions were generated by PCR-amplification in the presence of dNTPs containing either native C, 5mC, or 5hmC (Figure 1A). Successful incorporation of dCTP variants was confirmed by immunoblotting using specific antibodies against each variant (Figure 1B). The formation of R-loops during the in vitro transcription reactions was inspected by blotting immobilized RNAs with the S9.6 antibody (S9.6 Ab), which binds DNA:RNA hybrids (Figure 1C). To increase the specificity of hybrid detection, all samples were treated with RNase A in high salt conditions in order to digest all RNA molecules except those engaged in R-loops. The specific detection of DNA:RNA hybrids was confirmed by blotting transcription reaction products previously digested with RNase H (Figure 1C). In agreement with our hypothesis that 5hmC favours R-loops, increased amounts of DNA:RNA hybrids were detected in samples derived from in vitro transcription of 5hmC-rich ACTB P1, ACTB P2 and APOE DNA templates (Figure 1D). To exclude the possibility that our results were biased by an inherent preference of the S9.6 Ab for hybrids containing 5hmC, we performed electrophoretic mobility shift assays (EMSAs) using the S9.6 Ab and DNA:RNA hybrid substrates of the same sequence but containing C, 5mC, or 5hmC. The S9.6 Ab was able to delay the run of the three substrates with similar kinetics, indicating that the Ab equally recognizes DNA:RNA hybrids formed with any of the three C variants (Figure 1—figure supplement 1).
We then performed atomic force microscopy (AFM) to directly visualize R-loop structures obtained in the \emph{in vitro} transcription reactions (Figure 1E). R-loops were identified as previously described\textsuperscript{22,23}. Each individual DNA molecule establishing an R-loop structure in the AFM images was assigned manually. The frequency of these structures formed in the presence of C, 5hmC, or 5mC DNA templates was measured and normalized against the frequency formed in RNase H-treated samples (Figure 1E). In agreement with the hypothesis that transcription of 5hmC-rich DNA templates results in increased R-loop formation, AFM data revealed that R-loop structures are more frequently formed in the presence of 5hmC.

To investigate if the DNA modification impacts \emph{in vitro} transcription levels, we measured RNA synthesis from DNA templates containing unmodified C, 5hmC or 5mC (Figure 1—figure supplement 2). These data show that the T7 polymerase is highly sensitive to DNA modifications, since replacing C by either 5hmC or 5mC significantly decreased the transcript levels \emph{in vitro}. On one side, detecting more R-loops on a lower-transcription levels setting (i.e. 5hmC-rich templates) further strengthens our hypothesis that 5hmC increases R-loop formation. However, we cannot draw any conclusions regarding the impact of 5mC on R-loop formation, as a putative effect on R-loop levels could be masked by the significantly altered transcription. To clarify this aspect and further test our model, we continued our study with experiments performed \emph{in vivo}.

**TET enzymatic activity impacts endogenous R-loop levels**

To test whether the 5hmC DNA modification induces R-loop formation \emph{in vivo}, we quantified R-loop levels in mES cells after depletion of Tet1, Tet2 and Tet3. Despite the significant reduction in Tet enzymes (Figure 2—figure supplement 1A), the levels of 5hmC were not significantly affected by Tet1 or Tet2 depletion (Figure 2A and Figure 2—figure supplement 1B). In contrast, depletion of Tet3 resulted in a significant loss of 5hmC, an effect that was exacerbated by the simultaneous depletion of the three enzymes (Figure 2A and Figure 2—figure supplement 1B). No significant changes were observed in 5mC levels (Figure 2A and Figure 2—figure supplement 1B). This finding suggests that there is a partial redundancy in the activity of the three Tet enzymes in mES cells. The loss of Tet1 or Tet2 – but not of Tet3 - is compensated by the remaining Tets. In agreement with the hypothesis that 5hmC promotes R-
loop formation, dot-blot hybridization of total cellular nucleic acids using the S9.6 Ab revealed reduced endogenous R-loop levels in mES cells after depletion of Tet3 and after co-depletion of the three Tet enzymes (Figure 2B and Figure 2—figure supplement 1B). We also measured R-loop levels upon RNAi depletion of the Tet enzymes in NIH-3T3 mouse fibroblasts (Figure 2—figure supplement 2A). As observed in mES cells, a significant reduction of 5hmC, but not 5mC, was obtained upon depletion of Tet3 and of the three Tets in mouse fibroblasts (Figure 2—figure supplement 2B, C). The triple knockdown of the Tet enzymes significantly reduced the global levels of R-loops in mouse fibroblasts, whereas Tet3 depletion in these cells had a minor impact in R-loops (Figure 2—figure supplement 2B, D). This effect was further confirmed by measuring R-loops formed at selected active genes by DNA:RNA immunoprecipitation (DRIP) in mES cells (Figure 2C) and in mouse fibroblasts (Figure 2—figure supplement 2E). The DRIP assays confirmed that R-loops are less abundant upon depletion of Tet enzymes. Importantly, simultaneous depletion of the three enzymes did not affect the expression levels of the analysed genes in mES cells and in mouse fibroblasts (Figure 2D, Figure 2—figure supplement 2F). These data suggest that the activity of Tet enzymes promotes the formation of R-loops in absence of changes in transcription levels.

Next, we employed a modified CRISPR-based system to target TET enzymatic activity to specific loci24. We used a pool of three specific guide RNAs (gRNAs) to direct a catalytically inactive Cas9 nuclease fused to the catalytic domain of TET1 (dCas9-TET1) to the last exon of the APOE gene in human osteosarcoma (U-2 OS) cells. As a control, dCas9 was fused to an inactive mutant version of the TET1 catalytic domain (dCas9-dTET1). Local enrichment of 5hmC following dCas9-TET1 targeting at the APOE locus was confirmed by DNA immunoprecipitation using antibodies specific for 5mC or 5hmC modified nucleotides (Figure 2E). The highest levels of 5hmC were detected at the gene segment adjacent to the gRNAs-target region. R-loop levels detected by DRIP peaked significantly at the gRNAs-target and in the downstream region, upon tethering of dCas9-TET1 but not of dCas9-dTET1 (Figure 2F). These differences were not caused by changes in APOE gene expression levels (Figure 2G). The increased levels of R-loops detected far from the dCas9-TET1 target site are consistent with the view that R-loops have the capacity to extend from their inception locus. Accordingly, R-loops can be up to several hundred base pairs long, and may extend over the entire gene body of shorter and/or highly transcribed genes25,26. Collectively, these data suggest that editing 5hmC density by changing
the expression levels or the genomic distribution of TET enzymes influences R-loop formation in cells.

ShmC and R-loops overlap genome-wide at transcriptionally active genes

To further inspect the link between 5hmC and R-loops we performed computational analyses of 5hmC antibody-based DNA immunoprecipitation (hMeDIP-seq) and DNA:RNA immunoprecipitation (DRIP-seq) datasets from mES and HEK293 cells\(^5,27-29\). To assess individual genome-wide distribution profiles, R-loops density was probed over fixed windows of \(\pm 10\) kbp around the 5hmC peaks (Figure 3A and Figure 3 - figure supplement 1A). The resulting metagene plots and heatmaps revealed a marked overlap between 5hmC-rich loci and R-loops. This overlap is also evident in the individual distribution profiles of 5hmC and R-loops along two long regions of chromosome 17 (Figure 3 - figure supplement 2). Despite the distinct distribution patterns of 5hmC (well-defined peaks) and R-loops (reads spanning genomic regions with highly heterogeneous lengths, ranging between a few dozen to over 1 kb\(^5\)), we could obtain a statistically significant Pearson correlation coefficient between both (p<0.05) (Figure 3B and Figure 3 - figure supplement 1B). Furthermore, approximately half of all R-loops detected genome-wide occurred at 5hmC-containing loci (Figure 3C and Figure 3 - figure supplement 1C). Notably, we observed an overlap between 5hmC and R-loops in 6839 (51\%) out of the 13288 actively expressed genes (Figure 3D), a feature illustrated in the individual profiles of mouse and human genes (Figure 3E and Figure 3 - figure supplement 1D). Metagene profiles revealed very similar patterns of intragenic distribution, with both 5hmC and R-loops increasing towards the transcription termination site (TTS), where they reached maximum levels (Figure 3F). At the transcription start sites (TSS), however, the 5hmC DNA modification was mostly absent, whereas R-loops were abundant. The detection of R-loop peaks at TSS regions is in agreement with previous studies\(^4,9\) and imply that 5hmC is not necessary for co-transcriptional DNA:RNA hybridization and R-loop formation.

The observed overlap between 5hmC and R-loop peaks at the TTS raises the hypothesis that Tet activity may be involved in transcription termination by directing the formation of R-loops. Defects in transcription termination result in the accumulation of readthrough transcripts extending beyond the TTS\(^30\). In agreement with a role in transcription termination,
TET1-KO human ES cells displayed significantly higher levels of readthrough transcripts genome-wide, when compared to wt human ES cells (Figure 3G).

We then sought to simultaneously detect 5hmC and R-loops at the same loci in individual mES cells. We performed proximity ligation assays (PLA) using S9.6 and anti-5hmC antibodies (Figure 4A). Control reactions without primary antibodies and with each antibody alone did not produce a significant signal. Staining of mES cells with S9.6 and anti-5hmC antibodies gave rise to a robust PLA signal scattered throughout the nucleus, which was mostly lost after digestion of cells with RNase H (Figure 4B).

5hmC-rich loci are prone to DNA damage

Disruption of R-loop homeostasis is a well-described source of genomic instability\(^1\). For instance, co-transcriptional R-loops increase conflicts between transcription and replication machineries by creating an additional barrier to fork progression\(^{31,32}\). Such conflicts may cause DNA damage, including DSBs, which can be revealed using antibodies against γH2AX. Indeed, R-loops overlap with γH2AX-decorated chromatin at different locations such as TTS\(^33\). We then sought to investigate if 5hmC creates conditions for DNA damage by promoting R-loop formation. We analysed the genomic distribution of γH2AX by interrogating chromatin immunoprecipitation followed by sequencing (ChIP-seq) data from HEK293 cells\(^34\). The individual distribution profiles of γH2AX were analysed over fixed windows of ±10 kbp around the 5hmC peaks detected in the same cells (Figure 5A). The resulting metagene plots revealed marked enrichment of γH2AX at 5hmC-rich loci. The genic distribution of 5hmC and R-loops along three different genes further showed co-localization of the two marks with γH2AX (Figure 5B). Analysis of γH2AX and 5hmC distribution within active genes revealed a low yet statistically significant Pearson correlation coefficient (p<0.05) (Figure 5C).

R-loops formed at 5hmC-rich regions impact gene expression in mES cells

To gather insights into the functional impact of R-loops at 5hmC-rich DNA regions, we analysed whole-transcriptome (RNA-seq) of mES cells overexpressing RNase H, a condition resulting in genome-wide loss of R-loops\(^5\). Amongst the genes that were differentially expressed, we found that 64% and 48% of all downregulated and upregulated genes, respectively, displayed R-loops.
overlapping with 5hmC (Figure 6A). Pathway analysis revealed that these differentially expressed genes (Supplementary File 1) are involved in the mechanistic target of rapamycin (mTOR) (downregulated) and MYC (upregulated) signalling pathways (Figure 6B and C). mTOR and MYC are known to play opposite roles in establishing diapause, the temporary suspension of embryonic development driven by adverse environmental conditions\textsuperscript{35}, a stage that ES cells mimic when cultured in vitro. mTOR, a major nutrient sensor, acts as a rheostat during ES cell differentiation and reductions in mTOR activity trigger diapause\textsuperscript{36}. While overexpression of RNase H in mES cells did not reveal any significant changes in the cell cycle progression (Figure 6—figure supplement 1A, B), we observed a significantly decreased expression of genes related to pluripotency (Pou5f1) and germ layer commitment (Sox17, Sox6, Dll1) pathways (Figure 6D). These data support the view that R-loops formed upon TET epigenetic reprogramming regulate gene expression in stem cells.

DISCUSSION

In this study, we probed the hypothesis that 5hmC facilitates the co-transcriptional formation of non-canonical DNA secondary structures, known as R-loops. Data from in vitro transcription reactions and atomic force microscopy provide direct evidence showing that transcription through 5hmC-rich DNA favours R-loop formation. By depleting TET enzymes in mES cells and in fibroblasts, we demonstrate that TET activity increases cellular R-loop levels. Notably, the diminished levels of R-loops observed in TET-depleted cells did not result from impaired transcription, suggesting that 5hmC directly promotes R-loop formation. In agreement, tethering TET enzymes to a specific genomic locus using a CRISPR/Cas9-based system increases the levels of R-loops at the target locus.

As 5hmC is mostly absent from the TSS, other chromatin and DNA features (e.g. histone modifications, DNA-supercoiling or G-quadruplex structures\textsuperscript{1}) known to induce R-loop formation, are likely to operate in these regions. In contrast, the robust overlap between R-loops and 5hmC at the TTS of active genes suggest a putative causal link. Mechanistically, 5hmC may impact R-loop formation by either destabilizing the DNA duplex or by altering RNA polymerase II elongation rate. Indeed, 5hmC modifies the DNA helix structure by favouring
DNA-end breathing motion, diminishes the thermodynamic stability of the DNA duplex and destabilizes GC pairing\(^{17,21}\). It also weakens the interaction between DNA and nucleosomal histones\(^{17}\), which is thought to accelerate RNA polymerase II elongation but can also facilitate nascent RNA annealing with the template DNA strand favouring R-loop formation. Future studies will clarify which one of these mechanisms, if not all, contribute to the observed impact of 5hmC on R-loops.

Acting as a promoter of R-loops, well-established drivers of DNA damage\(^1\), 5hmC may indirectly harm genome integrity. Indeed, we found that 5hmC-rich loci are hotspots for DNA damage genome-wide. While such unscheduled R-loops formed at 5hmC-rich loci may threaten genomic integrity, regulated formation of R-loops at specific 5hmC-decorated loci may exert important regulatory roles. Indeed, R-loops play diverse physiological functions\(^1\), such as the regulation of gene expression. Our findings that genome-wide 5hmC and R-loops overlap robustly at the transcription termination site of active genes and that TET-deficiency drives transcription readthrough support a model whereby TET enzymes act upstream of R-loop formation during transcription termination\(^{37}\).

TETs may play dual roles as both oncogenic and tumour suppressor genes, with the former arising as the consequence of altered expression levels or function, as observed in several cancers, such as triple-negative breast cancer\(^{38,39}\). In addition to altering the expression levels of tumour suppressors or oncogenes\(^{38}\), our findings suggest that TET-driven changes in the DNA methylation landscape may as well drive transcription-dependent genome damaging events that could facilitate cancer development and progression. In agreement with this view, a TET1 isoform that lacks regulatory domains, including its DNA binding domain, but retains its catalytic activity, is enriched in cancer cells\(^{40}\), suggesting that mis-targeted TET activity may drive oncogenic events, such as genomic instability. Conversely, TET activity deposits 5hmC at DNA damage sites induced by aphidicolin or microirradiation in HeLa cells and prevents chromosome segregation defects in response to replication stress\(^{41}\). While the role that TETs play during carcinogenesis is not yet clear, the impact of 5hmC on stem cell differentiation and development has been extensively studied\(^{42}\). By driving the developmental DNA methylome reprogramming, TETs carry out numerous functions related to early developmental processes. Here, we disclose a putative new role for R-loops as mediators of 5hmC-driven gene expression programs in stem cells. Our gene ontology analysis revealed that R-loops formed at 5hmC-rich
regions impact the expression of genes involved in establishing diapause. This stage of temporary suspension of embryonic development is triggered by adverse environmental conditions. Accordingly, changes in the activity of mTOR, a major nutrient sensor, control ES cell commitment to trigger diapause. The mTOR signalling pathway was significantly downregulated upon global R-loop suppression by RNase H. Conversely, MYC targets, which prevent ES cells from entering the state of dormancy that characterizes diapause, were amongst the genes more significantly upregulated upon RNase H overexpression in mES cells. These cells also displayed reduced expression levels of genes related to pluripotency (Pou5f1) and germ layer commitment (Sox17, Sox6, Dll1) pathways. Whether the controlled 5hmC-driven formation of R-loops at specific genes drives stem cells fate and how do TET enzymes capture the environmental cues to target R-loop formation at selected genes are important questions that emerge from our findings. Our study sets the ground for further research aimed at investigating the role of R-loops in ES cells.

MATERIALS AND METHODS

Cell lines and culture conditions

E14TG2a (E14) mouse embryonic stem (mES) cells were provided by Domingos Henrique (Instituto de Medicina Molecular João Lobo Antunes), and were a gift from Austin Smith (Univ. of Exeter, UK). 129S4/SvJae (J1) mES cells were kindly provided by Joana Marques (Medical School, University of Porto). Cells were grown as monolayers on 0.1% gelatine (410875000, Acros Organics) coated dishes, using Glasgow Modified Eagle’s Medium (GMEM) (21710-025, Gibco), supplemented with 1% (v/v) 200mM L-glutamine (25030-024, Thermo Scientific), 1% (v/v) 100mM sodium pyruvate (11360-039, Gibco), 1% (v/v) 100x non-essential aminoacids solution (11140-035, Gibco), 0.1% (v/v) 0.1M 2-mercaptoethanol (M7522, Sigma Aldrich), 1% (v/v) penicillin-streptomycin solution (15070-063, Gibco) and 10% (v/v) heat-inactivated, ES-qualified FBS (SH30070, Cytiva). Medium was filtered through a 0,22μm filter. Home-produced leukaemia inhibitory factor (LIF) was added to the medium upon plating, at 6×10^-2 ng/μL. U-2 OS osteosarcoma, HEK293T embryonic kidney cells and NIH-3T3 mouse fibroblasts were purchased from ATCC. Cells were grown as monolayers in Dulbecco’s Modified Eagle medium (DMEM) (21969-035, Gibco), supplemented with 1% (v/v) 200mM L-glutamine (25030-024,
Thermo Scientific), 1% (v/v) penicillin-streptomycin solution (15070-063, Gibco) and 10% (v/v) FBS (10270106, Gibco). All cells were maintained at 37°C in a humidified atmosphere with 5% CO₂. Cell lines were authenticated using the STR profiling service provided by ATCC and routinely tested negative for mycoplasma contamination using the Mycoplasma Detection Kit (Invivogen, San Diego, CA).

**Tet knockdown**

For each Tet, a mixture of 4 siRNA provided as a single reagent was transfected using Lipofectamine RNAiMAX Transfection Reagent (13778150, Invitrogen) for 48h. All siRNAs were purchased as siGENOME SMARTPool from Dharmacon: mouse Tet1 (M-062861-01), mouse Tet2 (M-058965-01) and mouse Tet3 (M-054156-01). A siRNA targeting the firefly luciferase was used as control. For the Tet1/2/3 triple KD the three siRNA reagents were combined in the same RNA interference experiment. Tet3 knockdown was performed in J1 mES cells stably expressing a doxycycline-inducible short hairpin RNA targeting Tet3 (Supplementary File 2A). Cells were treated for 48h with 2 μg/mL doxycycline (D9891, Sigma Aldrich).

**RNA isolation and quantitative RT-PCR**

Total RNA was isolated using TRizol reagent (15596018, Invitrogen). cDNA was prepared through reverse transcriptase activity (MB125, NZYTech). RT-qPCR was performed in the ViiA 7 Real-Time PCR system (Applied Biosystems), using PowerUp SYBR Green Master Mix (A25918, Applied Biosystems). Relative RNA expression was estimated as follows: $2^{(\text{Ct reference} - \text{Ct sample})}$, where Ct reference and Ct sample are mean threshold cycles of RT-qPCR done in duplicate for U6 snRNA or Gapdh mRNA and for the gene of interest, respectively. Primer sequences are presented in Supplementary File 2B.

**Dot Blot of genomic R-loops, 5mC and 5hmC**

Cells were lysed in lysis buffer (100mM NaCl, 10mM Tris pH 8.0, 25mM EDTA pH 8.0, 0,5% SDS, 50 μg/mL Proteinase K) overnight at 37°C. Nucleic acids were extracted using standard phenol-chloroform extraction protocol and re-suspended in DNase/RNase-free water. Nucleic acids were then fragmented using a restriction enzyme cocktail (20U each of EcoRI, BamHI, HindIII, BsrGI and XhoI). Half of the sample was digested with 40U RNase H (MB085, NZYTech).
for 48h at 37°C, to be used as a negative control in R-loops blotting. Digested nucleic acids were cleaned with standard phenol-chloroform extraction and re-suspended in DNase/RNase-free water. Nucleic acids samples were quantified in a NanoDrop 2000 spectrophotometer (Thermo Scientific), and equal amounts of DNA were deposited into a positively charged nylon membrane (RPN203B, GE Healthcare). Membranes were UV-crosslinked using UV Stratalinker 2400 (Stratagene), blocked in 5% (m/v) milk in PBSt (PBS 1× containing 0.05% (v/v) Tween 20) for 1h at room temperature, and immunoblotted with specific antibodies. For the loading control, membranes were stripped in 0.5% SDS for 1h at 60°C, followed by blocking and re-probing. Details of antibodies used are included in Supplementary File 2C.

**Proximity Ligation Assay (PLA)**

E14 mES cells were grown on coverslips and fixed/permeabilized with methanol for 10min on ice, followed by 1min acetone on ice. Cells were then incubated with primary antibodies for 1h at 37°C, followed by a pre-mixed solution of PLA probe anti-mouse minus (DUO92004, Sigma Aldrich) and PLA probe anti-rabbit plus (DUO92002, Sigma Aldrich) for 1h at 37°C. Localized rolling circle amplification was performed using Detection Reagents Red (DUO92008, Sigma Aldrich), according to the manufacturer’s instructions. Slides were mounted in 1:1000 DAPI in Vectashield. For the RNase H control, fixed cells were treated with 3U/μL RNase H (MB085, NZYTech) for 1h at 37°C prior to incubation with the antibodies. Images were acquired using the Point Scanning Confocal Microscope Zeiss LSM 880, 63x/1.4 oil immersion, with stacking acquisition and generation of maximum intensity projection images. PLA foci per nucleus were quantified using ImageJ. Details of antibodies used are mentioned in Supplementary File 2C.

**g-blocks PCR**

Designed g-blocks were ordered from IDT (Supplementary File 2D), and PCR-amplified using Phusion High-Fidelity DNA Polymerase (M0530S, NEB), according to manufacturer’s instructions. M13 primers were used to amplify all fragments (Supplementary File 2B), in the presence of dNTP mixes containing native (MB08701, NZYTech), methylated (D1030, Zymo Research) or hydroxymethylated (D1040, Zymo Research) cytosines. Efficient incorporation of modified dCTPs was confirmed through immunoblotting with specific antibodies. Details of antibodies used are mentioned in Supplementary File 2C.
In vitro transcription

PCR products were subject to *in vitro* transcription using the HiScribe T7 High Yield RNA Synthesis Kit (E2040S, NEB), which relies on the T7 RNA polymerase to initiate transcription from a T7 promoter sequence (present in our fragments). Reactions were performed for 2h at 37°C, using 1 μg of DNA as template, according to manufacturer’s instructions. The resulting RNA was column-purified with NucleoSpin RNA isolation kit (740955.250, Macherey-Nagel) and quantified in a NanoDrop 2000 spectrophotometer (Thermo Scientific).

Dot blot of R-loops formed in *in vitro*

Half of each *in vitro* transcription product was treated with 10U RNase H (MB085, NZYTech) at 37°C overnight, to serve as negative control. Then, all samples were treated with 0,05U RNase A (10109142001, Roche) at 350mM salt concentration, for 15min at 37ºC, and ran on agarose gel. Nucleic acids were transferred overnight to a nylon membrane through capillary transfer. The membrane was then UV-crosslinked twice, blocked in 5% milk in PBS for 1h at room temperature, and incubated with the primary antibody at 4ºC overnight. Signal quantification was performed using ImageJ. Details of antibodies used are included in Supplementary File 2C.

Atomic Force Microscopy

RNase A-treated *in vitro* transcription products, treated or not with RNase H, were purified through phenol-chloroform extraction method and re-suspended in DNase/RNase-free water. DNA solution was diluted 1:10 in Sigma ultrapure water (with final 10mM MgCl₂) and briefly mixed to ensure even dispersal in solution. A 10μL droplet was deposited at the centre of a freshly cleaved mica disc, ensuring that the pipette tip did not contact the mica substrate. The solution was let to adsorb on mica surface for 1-2min to ensure adequate coverage. The mica surface was carefully rinsed with Sigma ultrapure water, so that excess of poorly bound DNA to mica is removed from the mica substrate. Afterwards, the mica substrate was dried under a gentle stream of argon gas for approximately 2min, making sure that any excess water is removed. DNA imaging was performed using a JPK Nanowizard IV atomic force microscope, mounted on a Zeiss Axiovert 200 inverted optical microscope. Measurements were carried out in tapping mode using commercially available ACT cantilevers (AppNano). After selecting a region of interest, the DNA was scanned in air, with scan rates between 0.5 and 0.9 Hz. The
setpoint selected was close to 0.3 V. Several images from different areas of the same sample were performed and at least three independent samples for each condition were imaged. All images were of 512 × 512 pixels and analysed with JPK data processing software.

**CRISPR-assisted SmC/5hmC genome editing**

Lentivirus containing dCas9-TET1 (#84475, Addgene) or dCas9-dTET1 (#84479, Addgene) coding plasmids, as well as one out of three gRNAs (gRNA_1, 2 and 3) coding plasmids designed for the APOE last exon, were produced in HEK293T cells co-transfected with the Δ8.9 and VSV-g plasmids using Lipofectamine 3000 Transfection Reagent (L3000015, Invitrogen). After 48h, cell culture supernatant was collected and filtered through a 0.45μm filter. Lentivirus were collected through ultracentrifugation (25000 rpm, 3h, 4°C) using a SW-41Ti rotor in a Beckman XL-90 ultracentrifuge. Virus were re-suspended in PBS 1× and stored at -80°C. For infection, a pool of lentivirus containing dCas9-TET1 or dCas9-dTET1, as well as gRNA_1, 2 or 3 coding plasmids, was used to infect seeded U-2 OS cells. After 24h, antibiotic selection was performed with 1.5 μg/mL puromycin, and infection proceeded for more 48h. 3 days post-infection, cells were harvested and genomic DNA was extracted for subsequent protocols.

**DNA:RNA Immunoprecipitation (DRIP)**

Cells were collected and lysed in 100mM NaCl, 10mM Tris pH 8.0, 25mM EDTA, 0.5% SDS, 50μg/mL Proteinase K overnight at 37°C. Nucleic acids were extracted using standard phenol-chloroform extraction protocol and re-suspended in DNase/RNase-free water. Nucleic acids were then fragmented using a restriction enzyme cocktail (20U each of EcoRI, BamHI, HindIII, BsrGI and Xhol), and 10% of the digested sample was kept aside to use later as input. Half of the remaining volume was digested with 40U RNase H (MB085, NZYTech) to serve as negative control, for 72h at 37°C. Digested nucleic acids were cleaned with standard phenol-chloroform extraction and re-suspended in DNase/RNase-free water. DNA:RNA hybrids were immunoprecipitated from total nucleic acids using 5μg of S9.6 antibody (MABE1095, Merck Millipore) in binding buffer (10mM Na₂HPO₄ pH 7.0, 140mM NaCl, 0.05% Triton X-100), overnight at 4°C. 50μl protein G magnetic beads (10004D, Invitrogen) were used to pull-down the immune complexes at 4°C for 2-3h. Isolated complexes were washed 5 times (for 1 min on
ice) with binding buffer and once with Tris-EDTA (TE) buffer (10mM Tris pH 8.1, 1mM EDTA). Elution was performed in two steps, for 15min at 55°C each, using elution buffer (50mM Tris pH 8.0, 10mM EDTA, 0.5% SDS, 60µg/mL Proteinase K). The relative occupancy of DNA:RNA hybrids was estimated by RT-qPCR as follows: \(2^{(\text{Ct Input}-\text{Ct IP})}\), where Ct Input and Ct IP are mean threshold cycles of RT-qPCR done in duplicate for input samples and specific immunoprecipitations, respectively. Data were normalized against the corresponding RNase H-treated samples, and plotted as absolute numbers or as fold change over control. Primer sequences are shown in Supplementary File 2B.

**5-(hydroxy)Methylated DNA Immunoprecipitation ((h)MeDIP)**

Cells were collected and lysed in 100mM NaCl, 10mM Tris pH 8.0, 25mM EDTA, 0.5% SDS, 50µg/mL Proteinase K overnight at 37°C. Samples were sonicated with 4 pulses of 15s at 10mA intensity using a Soniprep150 sonicator (keeping tubes for at least 1min on ice between pulses). Fragmented nucleic acids were cleaned with standard phenol-chloroform extraction protocol and re-suspended in DNase/RNase-free water. 10% of sample was kept aside to use later as input. The remaining volume was denatured by boiling the samples at 100°C for 10min, followed by immediate chilling on ice and quick spin. Samples were divided in half, and 5µg of anti-5mC antibody (61255, Active Motif) or 5µg of anti-5hmC antibody (39791, Active Motif) were used to immunoprecipitate 5mC and 5hmC, respectively, in binding buffer (10mM Na₂HPO₄ pH 7.0, 140mM NaCl, 0.05% Triton X-100), overnight at 4°C. 50µl protein G magnetic beads (10004D, Invitrogen) were used to pull-down the immune complexes at 4°C for 2-3h. Isolated complexes were washed 5 times (for 1 min on ice) with binding buffer and once with TE buffer (10mM Tris pH 8.1, 1mM EDTA). Elution was performed in two steps, for 15min at 55°C each, using elution buffer (50mM Tris pH 8.0, 10mM EDTA, 0.5% SDS, 60µg/mL Proteinase K). The relative occupancy of 5mC and 5hmC was estimated by RT-qPCR as follows: \(2^{(\text{Ct Input}-\text{Ct IP})}\), where Ct Input and Ct IP are mean threshold cycles of RT-qPCR done in duplicate for input samples and specific immunoprecipitations, respectively. Primer sequences are presented in Supplementary File 2B.

**Cell cycle analysis**
pEGFP-N1 (GFP coding plasmid used as control) was purchased from Addgene, and pEGFP-RNaseH1 (GFP-tagged RNase H coding plasmid) was kindly provided by Robert J. Crouch (NIH, USA). Seeded mES cells were transfected with GFP (control) or GFP-tagged RNase H coding plasmids using Lipofectamine 3000 Transfection Reagent (L3000015, Invitrogen). 24 or 48h later, cells were trypsinized and pelleted by centrifugation at 500×g for 5min. Cells were fixed in cold 1% PFA for 20min at 4°C, followed by permeabilization in 70% ethanol for 1h at 4°C. Cells were then treated with 25 μg/mL RNase A (10109142001, Roche) in PBS 1× at 37°C for 20min, followed by staining with 20 μg/mL propidium iodide (P4864, Sigma Aldrich) in PBS 1× for 10 min at 4°C. Flow cytometry was performed on a BD Accuri C6 (BD Biosciences) and data were analysed using FlowJo software.

**Electrophoretic Mobility Shift Assay (EMSA)**

DNA:RNA hybrids formed with either C-, 5hmC- or 5mC-containing DNA were obtained by incubating ssDNA with the complementary ssRNA in annealing buffer (100mM KAc, 30mM HEPES pH 7.5). Native and C-modified oligonucleotides were ordered from IDT (Supplementary File 2E). Hybrid formation was confirmed in a native polyacrylamide gel. Increasing amounts of S9.6 antibody (MABE1095, Merck Millipore) were added to the DNA:RNA hybrids and the complexes were ran in a native polyacrylamide gel to assess the S9.6 capacity to bind hybrids containing each of the three C variants. The amount of free probe was quantified using ImageJ.

**Multi-omics data**

High-throughput sequencing (HTS) data for mES cells and HEK293 cells were gathered from GEO archive: transcriptome of mES cells (GSE67583); R-loops in mES cells (GSE67581); 5hmC in mES cells (GSE31343); γH2AX in mES cells (GSE69140); active transcription in HEK293 (GRO-seq, GSE51633); R-loops in HEK293 (DRIP-seq, GSE68948); 5hmC modification in HEK293 (hMeDIP-seq, GSE44036); γH2AX (ChIP-seq, GSE75170). Transcriptome profiles of mES cells overexpressing RNase H were obtained from GSE67583. The quality of HTS data was assessed with FastQC (www.bioinformatics.babraham.ac.uk/projects/fastqc).

**5hmC, R-loop and γH2AX genome-wide characterization**

The HTS datasets produced by immunoprecipitation (DRIP-seq, ChIP-seq and hMeDIP-seq)
were analysed through the same workflow. First, the reads were aligned to the reference mouse and human genome (mm10 and GRCh38/hg38 assemblies, respectively) with Bowtie\textsuperscript{45}, and filtering for uniquely aligned reads. Enriched regions were identified relative to the input samples using MACS\textsuperscript{46}, with a false-discovery rate of 0.05. Finally, enriched regions were assigned to annotated genes, including a 4-kilobase region upstream the transcription start site and downstream the transcription termination site. Gene annotations were obtained from mouse and human Gencode annotations (M11 and v23 versions, respectively) and merged into a single transcript model per gene using BedTools\textsuperscript{47}. For individual and metaprofiles, uniquely mapped reads were extended in the 3’ direction to reach 150 nt with the Pyicos\textsuperscript{48}. Individual profiles were produced using a 20bp window. For the metaprofiles centered around 5hmC peaks: 5hmC enriched regions were aligned by the peak summit (maximum of the peak) and the read density for the flanking 10 kbp were averaged in a 200bp window. For the metagene profiles: the gene body region was scaled to 60 equally sized bins and ±10 kbp gene-flanking regions were averaged in 200bp windows. All profiles were plotted as normalized reads per kilobase per million mapped reads (RPKMs). A set of in-house scripts for data processing and graphical visualization were written in bash and in the R environmental language http://www.R-project.org\textsuperscript{49}. SAMtools\textsuperscript{50} and BEDtools were used for alignment manipulation, filtering steps, file format conversion and comparison of genomic features. Statistical significance of the overlap between 5hmC and R-loops was assessed with enriched regions and permutation analysis. Briefly, random 5hmC and R-loops enriched regions were generated 1000 times from annotated genes using the shuffle BEDtools function (maintaining the number and length of the originally datasets). The p-value was determined as the frequency of overlapping regions between the random datasets as extreme as the observed.

**Transcriptome analysis**

Expression levels (Transcripts per Million, TPMs) from RNA-seq and GRO-seq datasets were obtained using Kallisto\textsuperscript{51}, where reads were pseudo-aligned to mouse and human Gencode transcriptomes (M11 and v23, respectively). Transcriptionally active genes for 5hmC and R-loops annotation were defined as those with expression levels higher than the 25\textsuperscript{th} percentile. Differential expression in mES cells overexpressing RNase H was assessed using edgeR (v3.20.9) and limma (v3.34.9) R packages\textsuperscript{52,53}. Briefly, samples comparison was performed using voom transformed values, linear modelling and moderated T-test as implemented in limma R
package, selecting significantly differentially expressed genes with B-statistics higher than zero.

Significantly enriched pathways of up and down-regulated genes (with overlapping R-loops/5hmC regions) were selected using Fisher’s Exact Test and all expressed genes as background gene list. Evaluated pathways were obtained from the hallmark gene sets of Molecular Signatures Database (MSigDB)\textsuperscript{54} and filtered using False discovery rate corrected p-values < 0.05.

For the analysis of transcription readthrough, transcriptome profiles from human embryonic stem cells (WT and \textit{TET1} KO) were obtained from a GEO (GSE169209). RNA-seq data were mapped to the reference human genome (GRCh38) with the STAR v2.7.8a using default parameters\textsuperscript{55}. Transcription readthrough levels were evaluated by counting the number of reads mapping downstream the Transcription Termination Site (TTS) using ARTDeco\textsuperscript{56} and human genome annotation from GENCODE project (GENCODE release 37). Genes with an enrichment in transcriptional readthrough in \textit{TET1} KO samples relative to the control were identified. Metagene profiles were built using the \textit{computeMatrix} tool from the deepTools v3.5.1\textsuperscript{57} and default packages from Python language. Genes were scaled to equally sized bins of 100bp so that all annotated TSSs and TTSs were aligned. Regions of 1Kb were added upstream of TSS and downstream of TTS and also averaged in 100bp bins. All read counts were normalized by the number of mapped reads (RPKM).

**ACKNOWLEDGMENTS**

We thank our colleagues, Joana Marques, Domingos Henrique and Robert Crouch for kind gifts of cell lines, plasmids and reagents. We thank the technical support and resources provided by the Bioimaging and the Flow Cytometry Facilities of Instituto de Medicina Molecular João Lobo Antunes. This work was funded by PTDC/BIA-MOL/30438/2017 and PTDC/MED-OUT/4301/2020 from Fundação para a Ciência e Tecnologia (FCT), Portugal. Funding was also received from EU Horizon 2020 Research and Innovation Programme (RiboMed 857119). J.C.S. is the recipient of an FCT PhD fellowship PD/BD/128292/2017. Work in C.M.A.’s laboratory is supported by “la Caixa” Foundation and FCT, I.P. (LCF/PR/HP21/52310016; PTDC/BIA-MOL/6624/2020; PTDC/MED-ONC/7864/2020).
REFERENCES


41. Kafer, G. R. et al. 5-Hydroxymethylcytosine Marks Sites of DNA Damage and Promotes


---

**Figure Legends:**

**Figure1:** 5hmC favours co-transcriptional R-loop formation. (A) Native or modified dCTPs were incorporated upon PCR amplification into DNA fragments with sequences from the transcription termination region of *ACTB* (*ACTB P1* and *ACTB P2*) or *APOE*. (B) Incorporation of dCTP variants confirmed by immunoblotting using specific antibodies against 5mC, 5hmC and...
double-stranded DNA (dsDNA). (C) R-loops formed upon in vitro transcription reactions were detected by immunoblotting using the S9.6 antibody. RNase H-treated in vitro transcription reaction products (RH+) serve as negative controls. All data are representative of seven independent experiments with similar results. (D) S9.6 immunoblots were quantified and the R-loop levels normalized against the levels detected in the reaction products of DNA templates containing native C. Data represent the mean and standard deviation (SD) from seven independent experiments. *p<0.05, **p<0.01 and ***p<0.001, two-tailed Student’s t test. (E) In vitro transcription reaction products of ACTB P2 templates were visualized using atomic force microscopy. R-loop structures obtained from 5hmC-containing ACTB P2 transcription in the absence (RH-) or presence (RH+) of RNase H are shown. R-loops present in the transcription reaction products of C, 5mC or 5hmC-containing ACTB P2 templates were counted in a minimum of 80 filaments observed in three individual AFM experiments.

**Figure 2: TET enzymatic activity impacts R-loop levels.** Quantification of 5hmC and 5mC (A) and R-loops (B) dot blots of Tet1, Tet2, Tet3 single KD, and of Tet1/2/3 triple KD mES cells. Dot blots are shown in Figure 2 – figure supplement 1B. Data were normalized against dsDNA levels. *p<0.05, **p<0.01, ***p<0.001, two-tailed Student’s t test. ND=not detected. (C) R-loop levels assessed by DRIP, in Tet1/2/3 triple KD mES cells. Data were normalized against RNase H-treated samples. *p<0.05, **p<0.01, two-tailed Student’s t test. (D) Transcription levels of the genes presented in (C) assessed by RT-qPCR. 5hmC/5mC (E) and R-loop (F) levels determined by (h)MeDIP or DRIP at 4 regions of the APOE gene upon tethering of dCas9-TET1 or dCas9-dTET1 to the last exon of APOE in U-2 OS cells. R-loop data were normalized against RNase H-treated samples. *p<0.05, **p<0.01, two-tailed Student’s t test. (G) APOE transcription levels upon targeting dCas9-TET1 or dCas9-dTET1 to the last exon of the gene in U-2 OS cells. Data shown are the mean and SD from at least three independent experiments.

**Figure 3: 5hmC and R-loops overlap in active genes of mES cells.** (A) Metagene and heatmap profiles of 5hmC and R-loops probed over fixed windows of ±10 kbp around the 5hmC peaks in expressed genes. (B) Pearson correlation coefficient between 5hmC and R-loops distribution within active genes (p<0.05). (C) Number of loci displaying 5hmC, R-loops, and overlapping
5hmC and R-loops. *Permutation analysis, p<0.05. (D) Percentage of active genes displaying overlapping 5hmC and R-loops. (E) Individual profiles of 5hmC and R-loop distribution along the Acadl, Cr1l and Srr genes. Density signals are represented as reads per kilobase (RPKMs). (F) Metagene profiles of 5hmC and R-loops distribution in active genes. The gene body region was scaled to 60 equally-sized bins and ±10 kbp gene-flanking regions were averaged in 200bp windows. TSS: transcription start site. TTS: transcription termination site. Density signals are represented as RPKMs and error bars (gray) represent standard error of the mean. (G) Metagene profiles of genes showing transcription readthrough in wild-type and TET1 KO human ES cells. All gene regions were scaled to 2000 bp (gene body) and divided in equal bins of 100 bp. 1000 bp regions averaged in 100 bp bins were added upstream the TSS and downstream the TTS region. *p<0.05, Mann-Whitney rank test.

Figure 4: Simultaneous detection of 5hmC and R-loops at the same genomic loci in individual mES cells. (A) 5hmC and R-loops PLA foci in mES cells. DAPI was added to the mounting medium to stain DNA. Scale bars: 10μm. Data are representative of at least three independent experiments with similar results. (B) Boxplot showing 5hmC/R-loops PLA foci per nucleus. Horizontal solid lines represent the median values and whiskers correspond to the 10 and 90 percentiles. A minimum of 300 cells from at least three independent experiments was scored for each experimental condition. ****p<0.0001, Mann-Whitney rank test.

Figure 5: 5hmC-rich loci are genomic hotspots for DNA damage. (A) Metagene profiles of 5hmC and γH2AX probed over fixed windows of ±10 kbp around the 5hmC peaks in expressed genes of HEK293 cells. (B) Individual profiles of 5hmC, R-loops and γH2AX distribution along the EME2, HEXIM1 and RPL36 genes. Density signals are represented as reads per kilobase (RPKMs). (C) Pearson correlation coefficient between 5hmC and γH2AX at active genes (p<0.05).

Figure 6: Cellular pathways affected by R-loops formed at 5hmC loci. (A) Volcano plot displaying the differentially expressed genes in mES cells upon RNase H overexpression. Of all downregulated and upregulated genes, 64% and 48% displayed R-loops overlapping with 5hmC, respectively. (B-C) Pathway analysis of the genes that have R-loops overlapping with 5hmC and are differentially expressed upon RNase H overexpression. Shown are the
significantly downregulated (B) and upregulated (C) hallmark gene sets from MSigDB. False discovery rate (FDR), p<0.001. (D) Transcription levels of pluripotency and germ layer commitment genes in mES cells overexpressing RNase H. *p<0.05, two-tailed Student’s t test. Means and SDs are from five independent experiments.

Figure 1 – figure supplement 1. Cytosine modifications do not affect the detection of DNA:RNA hybrids by the S9.6 Ab (A) DNA:RNA hybrids formed through annealing of C, 5hmC or 5mC-containing ssDNA with the complementary ssRNA. Data are representative of three independent experiments. (B) EMSA of DNA:RNA hybrids incubated with increasing amounts of the S9.6 Ab in a native polyacrylamide gel. Data are representative of three independent experiments. (C) Quantification of free DNA:RNA hybrids. The graph shows means and SD from three independent experiments.

Figure 1 – figure supplement 2. In vitro transcription levels of different DNA templates. (A) RNA concentration obtained from in vitro transcription of DNA templates containing C, 5hmC, or 5mC. Data shown are the mean from three independent experiments. (B) Mean and SD of the RNA levels obtained in each reaction normalized to reactions using DNA templates with native cytosines. Data are from three independent experiments.

Figure 2 – figure supplement 1. Impact of Tet depletion in mES cells in R-loops, 5hmC and 5mC. (A) Tet1, Tet2 and Tet3 mRNA expression levels in mES cells 48h after RNAi depletion of Tet1, Tet2 or Tet3. *p<0.05, **p<0.01, ***p<0.001, two-tailed Student’s t test. Graph shows the mean and SD from three independent experiments. (B) Dot blot of R-loops, 5hmC and 5mC in Tet1, Tet2 and Tet3 single KD and in Tet1/2/3 triple KD mES cell extracts. dsDNA was detected after stripping and re-probing of the same membranes. Data are representative of a minimum of three independent experiments.
Figure 2 – figure supplement 2. Tet depletion impacts R-loop formation in mouse fibroblasts.

(A) Tet1, Tet2 and Tet3 mRNA expression levels in mouse fibroblasts 48h after RNAi depletion of Tet1, Tet2 or Tet3. **p<0.01, two-tailed Student’s t test. (B) Dot blot of R-loops, 5hmC and 5mC in Tet1, Tet2 and Tet3 single KD and in Tet1/2/3 triple KD mouse fibroblast extracts. dsDNA was detected after stripping and re-probing of the same membranes. Data are from three independent experiments. Quantification of the 5hmC, 5mC (C), and R-loops (D) dot blots shown in (B). *p<0.05, two-tailed Student’s t test. ND=not detected. (E) DRIPs performed in Tet1/2/3 triple KD mouse fibroblasts to assess the R-loop levels formed in the indicated genes. R-loop levels were normalized to the corresponding RNase H-treated sample. *p<0.05, **p<0.01, ***p<0.001, two-tailed Student’s t test. (F) Transcription levels of the genes presented in (E) in control and Tet1/2/3 triple KD mouse fibroblasts. All graphs show the mean and SD from a minimum of three independent experiments.

Figure 3 – figure supplement 1. Genome-wide analysis of 5hmC and R-loops in HEK293 cells.

(A) Metagene and heatmap profiles of 5hmC and R-loops probed over fixed windows ±10 kbp around the 5hmC peaks in expressed genes. (B) Pearson correlation coefficient between 5hmC and R-loops distribution within active genes (p<0.05). (C) Number of loci displaying 5hmC, R-loops, and overlapping 5hmC and R-loops. *Permutation analysis, p<0.05. (D) Individual profiles of 5hmC and R-loop distribution along the EME2, HEXIM1 and RPL36 genes. Density signals are represented as reads per kilobase (RPKM).

Figure 3 – figure supplement 2. Chromosomal distribution of R-loops and 5hmC. Individual profiles of 5hmC and R-loops along two long chromosomal regions. Density signals are represented as reads per kilobase (RPKMs) and were uploaded in UCSC Genome Browser (http://genome-euro.ucsc.edu/).

Figure 6 – figure supplement 1. Global R-loop suppression does not impact cell cycle progression of mES cells. (A) Flow cytometry analysis of propidium iodide-treated mES cells with ectopic expression of either GFP (control) or GFP-tagged RNase H for 24 or 48h. Data are
representative of five independent experiments. (B) Percentage of control and RNase H-overexpressing mES cells at each cell cycle stage. Means and SDs are from five independent experiments.

Source data figure legends:

Source data 1. Original, uncropped images of blots and gels.
Source data 2. Original, uncropped images of blots.

Supplementary Files:

Supplementary File 1. Differentially expressed genes upon RNase H overexpression.

A

DNA:RNA

C  5hmC  5mC  ssRNA  ssDNA

B

DNA:RNA

S9.6 (µg) 0 0.5 1 2 4 8 0 0.5 1 2 4 8 0 0.5 1 2 4 8

C

Free DNA:RNA hybrids

S9.6 Ab (µg) 0 2 4 6 8

C  5hmC  5mC
### DNA template vs RNA concentration

<table>
<thead>
<tr>
<th>DNA template</th>
<th>RNA concentration</th>
</tr>
</thead>
<tbody>
<tr>
<td>ACTB P1 C</td>
<td>414,40 ng/uL</td>
</tr>
<tr>
<td>ACTB P1 5hmC</td>
<td>271,00 ng/uL</td>
</tr>
<tr>
<td>ACTB P1 5mC</td>
<td>71,17 ng/uL</td>
</tr>
<tr>
<td>ACTB P2 C</td>
<td>128,77 ng/uL</td>
</tr>
<tr>
<td>ACTB P2 5hmC</td>
<td>40,30 ng/uL</td>
</tr>
<tr>
<td>ACTB P2 5mC</td>
<td>28,60 ng/uL</td>
</tr>
<tr>
<td>APOE C</td>
<td>188,63 ng/uL</td>
</tr>
<tr>
<td>APOE 5hmC</td>
<td>108,47 ng/uL</td>
</tr>
<tr>
<td>APOE 5mC</td>
<td>22,23 ng/uL</td>
</tr>
</tbody>
</table>

#### Bar Graph Analysis

- **ACTB P1**
  - C vs 5hmC: **p < 0.01**  
  - C vs 5mC: ***p < 0.001**  
- **ACTB P2**
  - C vs 5hmC: **p < 0.01**  
  - C vs 5mC: ***p < 0.001**  
- **APOE**
  - C vs 5hmC: **p < 0.01**  
  - C vs 5mC: ***p < 0.001**
Figure A: 5hmC and R-loops expression profiles.

Figure B: Scatter plot showing the correlation between 5hmC and R-loops expression levels.

Figure C: Venn diagram illustrating the overlap between 5hmC and R-loops expression sets.

Figure D: Pie chart depicting the distribution of active genes categorized by 5hmC/R-loops overlap.

Figure E: Heatmaps for different genes (Acad1, Cr11, Srr) showing expression levels of 5hmC and R-loops.

Figure F: Graphs illustrating the expression levels of 5hmC and R-loops across different genomic positions.

Figure G: Comparative graph showing the expression levels of R-loops in WT and TET1 KO conditions.