UNIVERSIDADE DE LISBOA
FACULDADE DE MEDICINA DE LISBOA

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

João Maria Rodrigues Perlico da Cruz Sabino

Orientador: Prof. Doutor Sérgio Alexandre Fernandes de Almeida

Tese especialmente elaborada para obtenção do grau de
Doutor em Ciências Biomédicas
Especialidade Biologia Celular e Molecular

2022

UNIVERSIDADE DE LISBOA

FACULDADE DE MEDICINA DE LISBOA

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

João Maria Rodrigues Perlico da Cruz Sabino

Orientador: Prof. Doutor Sérgio Alexandre Fernandes de Almeida
Tese especialmente elaborada para obtenção do grau de Doutor em Ciências Biomédicas
Especialidade Biologia Celular e Molecular

Júri:
Presidente: Doutor Mário Nuno Ramos de Almeida Ramirez, Professor Associado com Agregação e Vice-Presidente do Conselho Científico da Faculdade de Medicina da Universidade de Lisboa.
Vogais:
- Doctor Brian Luke, Heisenberg Professorship da Faculty of Biology, Johannes Gutenberg University, Mainz (Germany);
- Doutora Cristina Joana Moreira Marques, Investigadora Auxiliar da Faculdade de Medicina da Universidade do Porto;
- Doutora Joana Mendes Neves, Group Leader do Instituo de Medicina Molecular – João Lobo Antunes, unidade de investigação associada à Faculdade de Medicina da Universidade de Lisboa;
- Doutor João António Augusto Ferreira, Professor Associado da Faculdade de Medicina da Universidade de Lisboa;
- Doutor Sérgio Alexandre Fernandes de Almeida, Professor Associado com Agregação da Faculdade de Medicina da Universidade de Lisboa (Orientador);
- Doutor Domingos Manuel Pinto Henrique, Investigador Auxiliar da Faculdade de Medicina da Universidade de Lisboa.

Financiado por Fundação para a Ciência e a Tecnologia, no âmbito do Lisbon Biomedical and Clinical Research (LisbonBioMed) PhD Program

2022

A impressão desta tese foi aprovada pelo Conselho Científico da Faculdade de Medicina de Lisboa em reunião de 24 de maio de 2022

As opiniões expressas nesta publicação são da exclusiva responsabilidade do seu autor

Table of Contents

Acknowledgments

First of all, I want to thank my supervisor Prof. Sérgio Almeida. Thank you for accepting me first as a master student and later as a PhD candidate. Actually, it was my short masters internship that sparked a curiosity for scientific research and made me decide to pursue a PhD. Thank you for being always available, for the scientific discussions, for pushing my critical thinking, for the extraordinary guidance, for your trust and your so much needed optimism (towards myself and my work), and for your plain honesty. During the last 6 years I grew so much, both professionally and personally, and a lot of that I owe to you.

I thank the members of my PhD Thesis Committee, Prof. Vanessa Morais, Prof. Simão Rocha and Prof. Lars Jansen, for the scientific input and for all the questions they raised, which made me better prepared to discuss and defend my work.

To all my lab people, both former and current, I thank for the support and scientific discussions. Ram, Ana Duarte, Mafalda, Alexandra and Ana Rita, thank you for welcoming me in the lab, for the mentorship, for getting me started on the bench, and for contributing to this project. Cristiana and Eduardo, for your kind heart, and for always having a nice word and a smile to share, which I cherish so much. Madalena and Inês, thank you for showing me how work colleagues can become friends. You have always been by my side through every step of the way, in the ups and downs, lifting me up and brightening every moment (specially the most intense). You have helped me to see life in colors, and my scientific achievements are also, directly or indirectly, yours.

To my neighbors from Claus Azzalin Lab, thank you for all the shared knowledge (as well as reagents), for the scientific discussions and for the inputs in the lab meetings. Particularly, thank you Claus and my friend Patrícia for taking part in this project.

Ana, António, José, and all members of iMM Bioimaging facility, thank you for your support, for your patience and for your kind words.

To my colleagues and friends Mariana, Eunice, Elvira, Helena, Henrique and Nuno, thank you for sharing this journey with me, for the great moments, for caring. You have all contributed to make these years unforgettable and I look forward to share many more adventures with you.

And because a place is only as good as the people in it, I am deeply grateful for all the amazing people I have come across in iMM and made it such an extraordinary workplace for me. Thank you for the scientific discussions and for the vibrant and friendly environment. It was a pleasure to be part of this institution.

I am also thankful to my funding agency Fundação para a Ciência e a Tecnologia, and to LisbonBioMed PhD Program, for providing me with such an extraordinary opportunity.

To my friends, specially Rogério, Inês, Maria Beatriz, Bea, Joana, Mateus, thank you for your presence, for your support, for the conversations, for the fun, for showing me new perspectives of life and for helping me to keep my balance.

Por fim, quero agradecer à minha família. Aos meus pais, por serem o meu porto seguro e a minha maior certeza na vida, às minhas avós por todo o carinho, ao meu irmão pela preocupação disfarçada, e aos meus gatos pelo brilho nos meus olhos.

Obrigado.

Abbreviations and Acronyms

(h)MeDIP	5-(hydroxy)Methylated DNA Immunoprecipitation
4mC	4-methylcytosine
5caC	5-carboxylcytosine
5fC	5-formylcytosine
5hmC	5-hydroxymethylcytosine
5hmU	5-hydroxyuracil
5mC	5-methylcytosine
6mA	6-methyladenine
A	Adenine
ACTB	*β-actin*
AFM	Atomic force microscopy
AID	Activation-induced cytidine deaminase
ALS4	Amyotrophic lateral sclerosis 4
ALT	Alternative lengthening of telomeres
APOBEC	Apolipoprotein B mRNA editing enzyme complex
AQR	Aquarius
ATP	Adenosine triphosphate
BAMBI	BMP and activin membrane bound inhibitor
BER	Base excision repair
bp	Base-pairs
C	Cytosine
CGIs	CpG islands

ChIP	Chromatin immunoprecipitation
CPA	Cleavage and polyadenylation
CTCF	CCCTC-binding factor
DAPI	4′,6-diamidino-2-phenylindole
dCTPs	Cytosine deoxyribonucleotides
DDR	DNA damage response
DDX23	DEAD-box helicase 23
DHS	DNase-I-hypersensitive
DHX9	DEAH box protein 9
DNA	Deoxyribonucleic acids
DNMT	DNA methyltransferase
DRIP	DNA:RNA immunoprecipitation
DSBH	Double-stranded β-helix
DSBs	Double-strand breaks
dsDNA	Double-stranded DNA
DSIF	DRB-sensitivity-inducing factor
dsRNA	Double-stranded RNA
EMSAs	Electrophoretic mobility shift assays
ES	Embryonic stem
G	Guanine
GADD45A	Growth arrest and DNA damage inducible alpha
gRNAs	Guide RNAs
HATs	Histone acetyltransferases

HDACs	Histone deacetylases
HKMTs	Histone lysine methyltransferases
HP1	Heterochromatin protein 1
HR	Homologous recombination
IDAX	Inhibition of the Dv1 and Axin complex
KDM4A	Lysine-specific demethylase 4A
lncRNA	Long non-coding RNA
LSD1	Lysine-specific demethylase 1
MeCP2	Methyl-CpG binding protein 2
miRNA	Micro RNA
MLL1	Mixed-lineage leukemia 1
mRNA	Messenger RNA
mtDNA	Mitochondrial DNA
mTOR	Mechanistic target of rapamycin
ncRNA	Non-coding RNA
NELF	Negative elongation factor
OXPHOS	Oxidative phosphorylation
PAS	Polyadenylation site
PLA	Proximity ligation assay
Prc2	Polycomb repressive complex 2
PRMTs	Protein arginine methyltransferases
P-TEFb	Positive transcription elongation factor b
RER	Ribonucleotide excision repair

RITS	RNA-induced transcriptional silencing
RNA Pol II	RNA Polymerase II
RNA	Ribonucleic acids
RNAi	RNA interference
RNase H	Ribonuclease H
RNP	Ribonucleoprotein
RPKMs	Reads per kilobase per million mapped reads
S9.6 Ab	S9.6 antibody
SAM	S-adenosyl-L-methionine
SD	Standard deviation
SETX	Senataxin
siRNA	Small interfering RNA
SMUG1	Single-strand-selective monofunctional uracil DNA glycosylase 1
snRNA	Small nuclear RNA
SRSF1	Serine/arginine splicing factor 1
SSBs	Single-strand breaks
ssDNA	Single-stranded DNA
ssRNA	Single-stranded RNA
T	Thymine
TARID	TCF21 antisense RNA inducing promoter demethylation
TCF21	Transcription factor 21
TDG	Thymine DNA glycosylase
TERRA	Telomeric-repeat-containing RNA

TET	Ten-eleven translocation
TGF-β	Transforming growth factor β
THO/TREX	Transcription and Export complex
THOC1	THO Complex 1
TPMs	Transcripts per Million
TSSs	Transcription start sites
TTS	Transcription termination site
UHRF1	Ubiquitin-Like-Containing PHD And RING Finger Domains 1
UTRs	Untranslated regions
VIM	Vimentin
XIST	X (inactive)-specific transcript
XRN2	5'-3' exoribonuclease 2

Resumo

A transcrição génica é um processo inerentemente mutagénico que requer mecanismos de vigilância rigorosos de forma a garantir a integridade genómica. Durante a transcrição, a molécula de RNA nascente pode hibridar com a cadeia complementar de DNA, dando origem a um híbrido de DNA:RNA e deixando assim a cadeia de DNA codificante desemparelhada. Estas estruturas de cadeia tripla são designadas R-loops e atuam como intermediários com relevância fisiológica em vários processes celulares, tais como a recombinação da classe das imunoglobulinas ou a expressão génica. No entanto, a sua formação descontrolada e persistente constitui uma importante fonte de lesões no DNA, nomeadamente quebras na dupla cadeia de DNA (*do inglês* DSBs). Para preservar a integridade do genoma, as células possuem diversos mecanismos para impedir a formação de R-loops ou para os remover. Topoisomerases e proteínas de ligação ao RNA limitam a formação de R-loops, enquanto helicases e ribonucleases, como RNase H1 e RNase H2, degradam-nos através da digestão da cadeia de RNA do híbrido. A ação articulada destas enzimas em diferentes fases do ciclo de transcrição e em contextos fisiológicos distintos é crucial para manter a homeostase da expressão génica e prevenir lesões de DNA decorrentes do processo de transcrição.

Características intrínsecas ao DNA transcrito influenciam a propensão para formar R-loops. Uma distribuição assimétrica de nucleótidos de guanina (G) e citosina (C) na dupla cadeia de DNA, com excesso de Cs na cadeia não codificante (enviesamento G:C positivo), favorece a formação de R-loops. Quando isso acontece, a cadeia codificante desemparelhada, rica em Gs, pode adquirir conformações quadruplas de guaninas (*G-quadruplexes*) que contribuem para a estabilização dos R-loops. Adicionalmente, a torção negativa que se acumula no DNA a montante de uma RNA Polimerase II (RNA Pol II) em transcrição leva a uma relaxação local do DNA, o que cria oportunidade para o RNA nascente invadir a dupla hélice e hibridar com o DNA complementar, promovendo assim a ocorrência de R-loops.

Por outro lado, os R-loops também podem causar modificações na cromatina, fundamentais à regulação da transcrição. R-loops próximos de promotores potenciam o recrutamento do complexo de acetiltransferases de histonas Tip60-p400, enquanto simultaneamente inibem a ligação de complexos supressores de transcrição e impedem a metilação da lisina 27 da histona H3. Quanto às regiões de terminação, ricas em Gs, a

formação de R-loops promove a dimetilação da lisina 9 da histona H3, uma modificação repressora que impõe o abrandamento da RNA Pol II durante a terminação da transcrição.

Além de afetarem as modificações de histonas, os R-loops atuam como barreiras à propagação da metilação em genes ativos. A metilação do DNA, na forma de 5-metilcitosina (5mC), resulta da junção covalente de um grupo metilo ao carbono 5 de uma citosina unida a uma guanina por uma ligação fosfodiéster (CpG). Esta reação é catalisada pelas enzimas metiltransferases de DNA (DNMT), que difundem 5mC por todo o genoma dos mamíferos, onde esta modificação desempenha papéis importantes ao nível da supressão de retrotransposões, silenciamento da transcrição e regulação global da expressão génica. Por exemplo, mais de 70% dos promotores de genes humanos contêm extensões de dinucleótidos CpG, denominadas ilhas CpG (CGIs), cuja atividade transcricional é suprimida pela metilação.

R-loops localizados junto a promotores de genes ativos mantêm as CGIs não metiladas, possivelmente pela redução da afinidade das enzimas DNMT ao DNA, ou pelo recrutamento de maquinaria específica de demetilação: as dioxigenases de metilcitosina designadas "translocação dez-onze" (*do inglês* TET). Em mamíferos, a família TET inclui as TET1, TET2 e TET3, que partilham a capacidade de oxidar 5mC em 5-hidroximetilcitosina (5hmC), uma modificação de DNA relativamente rara e muito menos frequente no genoma em comparação com 5mC. As TET conseguem ainda oxidar sequencialmente 5hmC em 5-formilcitosina (5fC) e 5-carboxilcitosina (5caC), modificações que podem seguir diversas vias com vista à sua reposição por citosina nativa. A nível genómico, verifica-se uma maior abundância de 5hmC em regiões regulatórias, tais como promotores e exões, em linha com a sua função na regulação da expressão génica. Promotores ativos apresentam enriquecimento em 5hmC, observado por exemplo aquando da ativação de programas de transcrição específicos em neurónios e progenitores neurais.

Curiosamente, 5hmC tem o potencial de alterar a estrutura da hélice de DNA, favorecendo atributos dinâmicos que lhe conferem maior acessibilidade. A interação entre o DNA e o dímero de histonas nucleossomal H2A-H2B é enfraquecida por 5hmC, facilitando o desmantelamento transiente do nucleossoma necessário à passagem da RNA Pol II ao longo da transcrição. Além disso, 5hmC diminui a estabilidade termodinâmica da cadeia dupla de DNA: enquanto 5mC aumenta a temperatura de desnaturação do DNA, 5hmC reduz a quantidade de energia necessária para separar as duas cadeias. Esta modificação também tem implicações ao nível do emparelhamento de bases, como observado através de simulações de dinâmicas moleculares que revelaram uma maior amplitude de flutuações

entre pares GC na presença de 5hmC, enquanto 5mC produz as amplitudes de flutuação mais baixas. 5hmC destabiliza os pares GC porque alivia limitações conformacionais através de um aumento da polaridade molecular, tornando assim o DNA menos rígido e inflexível.

Uma vez que características que destabilizam a dupla hélice de DNA, tais como torções ou *G-quadruplexes*, facilitam a hibridação da molécula de RNA nascente com a cadeia de DNA complementar, colocámos a hipótese de que a modificação 5hmC possa também favorecer a formação de R-loops. Para testar esta possibilidade, utilizámos primariamente um modelo de transcrição *in vitro*, que nos permitiu concluir que a presença de 5hmC no DNA transcrito promove de facto a hibridação do transcrito produzido com o DNA, criando R-loops.

O estudo desta hipótese foi então aprofundado *in vivo.* Para tal, alterámos a densidade de 5hmC, quer de forma generalizada através de variações nos níveis de expressão das TET, quer de forma focalizada através do direcionamento específico da sua ação enzimática, investigando posteriormente o consequente impacto nos R-loops. A depleção das três TET (e resultante redução significativa dos níveis celulares de 5hmC) causou um decréscimo acentuado de R-loops endógenos em células estaminais embrionárias e em fibroblastos de ratinho. A depleção de cada TET individualmente produziu resultados muito mais ligeiros, sugerindo a existência de uma redundância parcial na atividade das três enzimas. Adicionalmente, com recurso ao sistema CRISPR, direcionámos a actividade enzimática TET para um gene ativo específico, onde observámos aumento da ocorrência de R-loops. De salientar que os efeitos acima descritos ocorreram na ausência de alterações nos níveis de transcrição, cimentando um efeito direto da modificação 5hmC como facilitadora de híbridos DNA:RNA.

A interação entre 5hmC e R-loops foi corroborada por uma análise genómica global que revelou uma forte sobreposição, verificada em metade dos genes ativos, entre ambas as estruturas, sobreposição essa que validámos através de técnicas de co-localização de moléculas individuais. A distribuição de 5hmC e R-loops ao longo do perfil dos genes mostrou que o seu pico de sobreposição ocorre na zona de terminação da transcrição (*do inglês* TTS), sugerindo que nesta região as TET desempenham um papel específico conducente à formação de R-loops. De facto, em células com atividade deficitária das TET, observámos um aumento significativo de transcritos que se prolongam além do TTS, chamados *readthrough transcripts*, característicos de um processo de terminação erróneo. Estas evidências suportam um modelo segundo o qual a ação das TET induz a formação de R-loops necessários à terminação eficiente da transcrição. Além disso, constatámos ainda

que existe uma correlação positiva entre regiões ricas em 5hmC e marcadores de cromatina característicos da resposta a lesões no DNA. Dado o efeito de 5hmC como facilitador de R-loops, propomos que locais do genoma abundantes em 5hmC marcam focos de instabilidade genómica.

Finalmente, quisemos explorar o impacto funcional de R-loops formados em zonas abundantes em 5hmC. Uma vez que as TET impulsionam a reprogramação do metiloma subjacente ao desenvolvimento embrionário, a marca epigenética 5hmC tem um papel amplamente reconhecido na diferenciação de células estaminais. Como tal, decidimos utilizar dados de transcritómica de células estaminais embrionárias com supressão global de R-loops. A análise da expressão génica revelou que a depleção de R-loops posicionados especificamente em locais ricos em 5hmC influencia mais significativamente vias celulares relacionadas com o controlo do equilíbrio entre proliferação e dormência de células estaminais. Assim, este estudo revela um potencial papel dos R-loops, instruídos pela deposição controlada de 5hmC, como mediadores da activação de programas de transcrição específicos em células estaminais.

Summary

Transcription is an inherently mutagenic process that requires tight surveillance mechanisms to guarantee the preservation of genomic integrity. During transcription, the nascent RNA molecule can hybridize with the template DNA and form a DNA:RNA hybrid, displacing the single-stranded DNA (ssDNA). Although these triple-stranded structures, called R-loops, are physiologically relevant intermediates of several cellular processes, such as immunoglobulin class-switch recombination and gene expression, 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 to resolve them. R-loop formation is restricted by RNA-binding proteins and topoisomerases, whereas R-loops are removed by helicases and ribonucleases, such as ribonuclease H enzymes RNase H1 and RNase H2, which degrade R-loops by digesting the RNA strand of the DNA:RNA hybrid. The concerted action of several R-loop resolving enzymes at different stages of the transcription cycle and in distinct physiological contexts is extremely important to maintain gene expression homeostasis and to prevent transcription-dependent DNA damage.

Intrinsic features of the transcribed DNA influence the propensity to form R-loops. An asymmetrical distribution of guanines (G) and cytosines (C) nucleotides in the DNA duplex, with an excess of Cs in the template DNA strand (positive G:C skew), favors R-loop formation, which is further stabilized by the establishment of G quadruplexes in the G-rich coding strand. Additionally, the negative DNA supercoiling accumulating upstream of a transcribing RNA Polymerase II (RNA Pol II) creates a local DNA unwinding that provides a window of opportunity for nascent RNA to hybridize with the template DNA, hence promoting R-loop formation.

R-loops can drive chromatin modifications that are critical for transcription regulation. 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. Also, R-loops formed over G-rich terminator elements promote histone H3 lysine-9 dimethylation, a repressive mark that reinforces RNA Pol II pausing during transcription termination.

Besides affecting histone modifications, R-loops 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, retrotransposon silencing, and gene expression regulation. 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.

R-loops positioned near promoters of active genes maintain CGIs in an unmethylated state, likely by reducing the affinity of DNMT1 binding to DNA, or by recruiting active DNA demethylation machinery: the ten-eleven translocation (TET) methylcytosine dioxygenases. In mammals, TET family comprises TET1, TET2 and TET3, which share the ability to oxidize 5mC to 5-hydroxymethylcytosine (5hmC), a relatively rare DNA modification found across the genome much less frequently than 5mC. TETs can further oxidize 5hmC to 5-formylcytosine (5fC) and to 5-carboxylcytosine (5caC), which engage in different pathways that lead to their replacement by native cytosine. Genome-wide, 5hmC is more abundant at regulatory regions such as promoters and exons, consistent with its role in gene expression regulation. The levels of 5hmC are enriched at active promoter regions, as observed upon activation of neuronal function-related genes in neural progenitors and neurons.

Interestingly, 5hmC has the potential to modify the DNA helix structure by favoring DNA-end breathing motion, a dynamic feature of the protein–DNA complexes thought to control DNA accessibility. Moreover, 5hmC weakens the interaction between DNA and nucleosomal H2A-H2B dimers, facilitating transient nucleosome unfolding to accommodate the passage of RNA Pol II during transcription elongation. Also, 5hmC diminishes the thermodynamic stability of the DNA duplex: while 5mC increases the melting temperature, 5hmC reduces the amount of energy needed to separate the two strands of the DNA double helix. 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 with the lower amplitude values. 5hmC destabilizes GC pairing by alleviating steric constraints through an increase in molecular polarity, rendering the DNA more flexible and less rigid.

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 favor R-loop formation. We used an *in vitro* transcription model to show that the presence of 5hmC in the transcribed DNA promotes the annealing of the nascent RNA to the template DNA strand, leading to the formation of an R-loop.

This hypothesis was further tested *in vivo* by editing 5hmC density and assessing the consequent impact on R-loops. Indeed, depletion of the three TET enzymes, which significantly reduced cellular 5hmC levels, caused a pronounced decrease in endogenous R-loops in mouse embryonic stem (ES) and fibroblast cells. Interestingly, the results obtained with individual depletion of each single TET were much milder, suggesting that there is a partial redundancy in the activity of the three TET enzymes. Additionally, CRISPR-mediated tethering of TET to an active gene promoted the formation of R-loops. The above described effects occurred independently of changes in transcription rate, firmly pointing towards a direct impact of 5hmC on DNA:RNA annealing. 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.

The interplay between 5hmC and R-loops was further strengthened by genome-wide analysis revealing a strong overlap, detected in half of all active genes, between both structures, which was validated through single-molecule co-localization techniques. Metagene plots of 5hmC and R-loops density show that overlapping of 5hmC and R-loops peaks at the transcription termination site (TTS), suggesting a dedicated role of TET activity in transcription termination by guiding the formation of R-loops. Strikingly, TET abrogation leads to significantly higher levels of readthrough transcripts genome-wide, a characteristic of defective termination. These data support a model whereby TET enzymes act upstream of R-loop formation during efficient transcription termination. Owing to the effect of 5hmC as R-loop facilitator, we also described a positive correlation between 5hmC-rich regions and DNA damage response markers, positioning 5hmC decorated loci as genomic instability hotspots.

Finally, we wanted to explore the functional impact of R-loops formed in 5hmC-rich regions. Since TETs drive the developmental DNA methylome reprogramming, the role of 5hmC on stem cell differentiation and development has been widely acknowledged. As such, we used transcriptomic data from ES cells with global R-loop suppression. Gene expression analysis revealed that depletion of R-loops specifically positioned in 5hmC-rich loci impinges most significantly on pathways that control the proliferation/ dormancy balance in ES cells. Therefore, this study discloses a putative role for R-loops, instructed by controlled 5hmC deposition, as mediators in the activation of ES cells gene expression programs.

1. Introduction

1.1. From genetics to epigenetics – an overview

Civilization has soon acknowledged the influence of heredity and its evident effects on crops cultivation, domestic animals breeding, or even human family resemblance. But it was not until the 19th century, when Mendel's work demonstrated the first evidence of genetic inheritance, providing the mathematical foundation for the science of genetics, that heredity mechanisms started to be understood. Mendel's hypotheses were later reinforced by the establishment of genes as basic units, that are transmitted to offspring, and shape the reactions and processes occurring within the cell. Then, in approximately 50 years, our comprehension of genetics vastly increased, as the fields of biochemistry and molecular biology underwent major breakthroughs. Among other discoveries, chromosomal organization of the genome was drafted, deoxyribonucleic acid (DNA) was identified as the biochemical structure that composes our genetic material, the adenine-thymine and guanine-cytosine pairing was discovered, the DNA double helix structure was revealed and the human genome was sequenced[1,2].

For some decades, the genes were perceived as classical units of function, transmission, recombination and mutation. However, a deeper knowledge of molecular biology influenced gene definition criteria. The discovery of gene regulatory regions challenged commonly established gene boundaries, non-coding ribonucleic acids (RNA) changed the standard understanding of protein-coding sequences, and even alternative splicing or gene assembling phenomenon, observed for instance in immunogenetics, defied the "one gene – one protein" dogma[3].

Interestingly, this shift from the classical to the molecular perspective of genetics also raised discussion about one other major concept in the field: epigenetics. Introduced in 1942 by the embryologist Conrad Waddington, the term was used to designate the so far unknown developmental processes through which a fertilized zygote could evolve into a mature organism, with cells of varied phenotype. In other words, it refers to the complex pathways that connect genotype to phenotype. This concept was popularized in 1953 by Waddington's model of an "epigenetic landscape" (Figure 1), illustrating several developmental trails a cell may take during differentiation. The topology of the landscape, which will influence cell fate, is shaped by the genes below it[4].

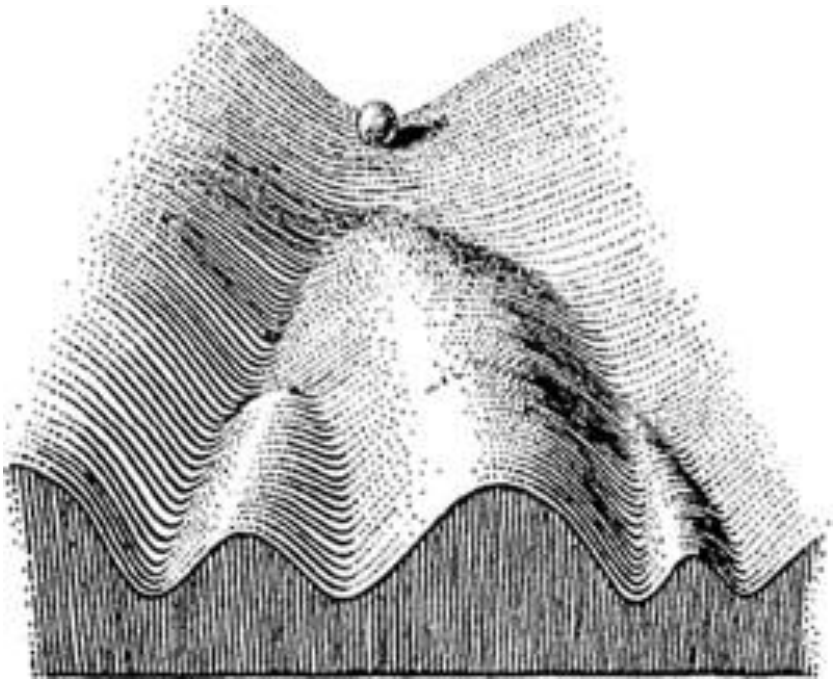


Figure 1: Picture of the epigenetic landscape proposed by Conrad Waddington. The model illustrates the multiple pathways a cell may take during the course of differentiation. Adapted from Baedke, 2013[172].

However, the growing understanding of gene regulation in eukaryotes highlighted the highly variable gene expression profiles among one organism's somatic cells, even though they all carry the same genomic information, and that these patterns can be clonally inherited. Therefore, the concept of epigenetics was adapted to cover this sequence-independent gene regulatory layer. Nowadays, although its definition is still under debate, epigenetics is commonly considered "the study of changes in gene function that are mitotically and/ or meiotically heritable and that do not entail a change in DNA sequence"[5]. Thus, the epigenome emerges as an extremely important field of study, that bridges the genome with (healthy) phenotype development. Epigenome disruption, either by inheritance or due to environment constraints[6], can lead to a myriad of severe pathologies, such as cancer, chromosomal instability syndromes, and neurological dysfunctions. Thus, the epigenetic balance requires tight regulation and fine-tuning[7].

1.2. Epigenetic systems

Several epigenetic systems have been described over the years. They compose dynamic layers of gene expression regulation, enabling cellular response to developmental and environmental cues.

1.2.1. Non-coding RNAs

Protein-coding genes represent only 2% of the total human genome. The transcriptional output of the extensively conserved genome regions that do not encode for proteins, previously thought as "junk" DNA, consist of non-coding RNAs (ncRNAs), which in fact play key roles in a variety of cellular functions, such as transcription, DNA replication, messenger RNA (mRNA) stability and processing, etc [8]. ncRNAs are found in viruses, bacteria, and eukaryotes, and can be categorized according to their length and physiological function[9]. Some examples are: (1) micro RNAs (miRNAs), which are short regulatory RNAs with a wide range of functions, including developmental processes related to pregnancy (such as placental development and the control of maternal-fetal immunological balance), host–microorganism interactions (as the case of viruses' miRNAs that interact with insect hosts to help infection), or even in gene silencing through mRNA cleavage[9]; (2) small interfering RNAs (siRNAs), that also lead to specific post-transcriptional gene silencing via mRNA degradation[9]; (3) small nuclear RNAs (snRNAs) which, among other functions, play an important role in splicing events, as part of the spliceosome complex that is assembled around newly transcribed pre-mRNA[8]; (4) long non-coding RNAs (lncRNAs), composed of spliced, capped and polyadenylated transcripts with over 200 base-pairs (bp), that may associate directly with proteins and mRNA and cause chromatin changes that lead to gene activation or repression. A well-described example is the X (inactive)-specific transcript (*XIST*) lncRNA, involved in the transcriptional silencing of one of the X chromosomes in female mammals. Some lncRNAs are encoded on the antisense strand of a protein-coding gene, or even within an intron sequence, regulating the expression of the respective gene[8,9].

1.2.2. Histone modifications

The eukaryotic genome is mainly organized in a highly compact structure in the nuclei that relies on the nucleosome as its repetitive unit. Each nucleosome consists of 147

bp of DNA wrapped around a core histone octamer, composed of two of each H2A, H2B, H3 and H4 histones. Structure-wise, the octamer has a central $(H3\text{-}H4)_2$ tetramer, flanked on each side by H2A-H2B dimers. H1, the linker histone, binds to the entry and exit sites of DNA on the surface of the nucleosome[10]. First reported in 1964, histones can undergo post-translational modifications[11], which alter the dynamic structure of chromatin and recruit specific remodelling proteins and enzymes. Therefore, histone modifications impact on DNA-based processes, such as gene transcription, DNA replication and repair[12]. The more common histone modifications are acetylation, methylation and phosphorylation.

Histone acetylation consists in the transfer of an acetyl group to the lysine residues within the N-terminal tail protruding from the nucleosome core. Histone acetyltransferases (HATs) mediate this transfer, while acetyl removal is carried by histone deacetylases (HDACs). Acetylation neutralizes lysine's positive charge, therefore weakening the interaction between histones and the negatively charged DNA, which causes a relaxation of chromatin that fosters transcription. Removal of the acetyl group stabilizes the nucleosome, positioning HDACs as transcription repressors[4,12].

Histone phosphorylation usually occurs on serine, threonine and tyrosine residues, predominantly in the N-terminal histone tails, although phosphorylation sites also exist in the nucleosome core. Kinases transfer a phosphate group from adenosine triphosphate (ATP) to histones, adding negative charge that influences chromatin architecture. Phosphatases remove the phosphate group[12]. In mammals, the most notorious case of histone phosphorylation happens during DNA damage response (DDR), as phosphorylation of the histone H2A(X) (H2A variant) at serine 139 occurs rapid and abundantly to signal DNA break sites, creating γH2AX decorated loci[13]. Histone phosphorylation is also connected to gene expression regulation. On one hand, the association of H3 phosphorylation with acetylation marks involves this modification in chromatin relaxation and transcription activation. On the other hand, during mitosis and meiosis, H3 phosphorylation (namely H3S10P) is linked to chromosome compaction and segregation through the establishment of condensed chromatin states[13].

Histone methylation occurs mainly on lysines and arginines, and it does not change histones' charge. Histone lysine methyltransferases (HKMTs) catalyse fairly specific reactions: they either transfer one, two, or three methyl groups to specific substrates and lysine residues. Methylation removal is performed by lysine demethylases, which also demonstrate high substrate specificity, as well as sensitivity to histone methylation degree. Examples are the lysine-specific demethylase 1 (LSD1) or the lysine-specific demethylase

4A (KDM4A)[12]. In turn, methylation of arginines is catalysed by a family of protein arginine methyltransferases (PRMTs), which act on a variety of substrates and induce mono or dimethylation. Demethylation of arginines remains more elusive. Nevertheless, there is evidence of jumonji protein JMJD6's capacity to demethylate arginine[12,14]. Histone methylation exerts prominent effects on gene expression regulation. For instance, H3K4me3 and H3K4me1 are associated with active transcription and promoter activity, whereas H3K27me3 is a marker of repressive chromatin. Moreover, histone marks may have context-dependent functions, and both collaborative and antagonistic relationships between different histone modifications have been reported[14].

1.2.3. DNA methylation

DNA methylation is the most prominent epigenetic mark in the mammalian genome, and it occurs through the covalent addition of a methyl group (CH_3) to the carbon 5 of a cytosine base, generating 5-methylcytosine (5mC)[15]. In bacteria, also N6-methyladenine (6mA) and N4-methylcytosine (4mC) DNA methylation modifications have been reported. The presence of 6mA in mammal DNA is controversial, as evidence suggests it might remain undetected due to the sensitivity limitations of detection techniques[16]. On the RNA, however, 6mA is the most prevalent internal modification. It is enriched in the mammalian nervous system, where it plays a role in mRNA metabolism regulation, consequently affecting brain function, neuronal development and neurological disorders[17].

DNA methylation dynamics comprise methylation establishment, maintenance and demethylation. The transfer of a methyl group to DNA is catalysed by the DNA methyltransferase (DNMT) family of enzymes, which are present in several organisms, from bacteria to humans. S-adenosyl-L-methionine (SAM) is the methyl donor[15]. In mammals, this family is composed of DNMT1, DNMT3A, DNMT3B and the cofactor DNMT3L. DNMT3A and DNMT3B are responsible for *de novo* DNA methylation. They contain a highly conserved DNMT domain (MTase domain) in the carboxy-terminus, as well as two chromatin reading domains. DNMT3L is a catalytically inactive DNMT known to interact and stimulate the activity of DNMT3A and DNMT3B specifically in the germline[18]. Recently, DNMT3C has been described as another *de novo* methyltransferase that evolved in rodents through the duplication of the *Dnmt3B* gene, with a specific role in the epigenetic control of retrotransposons[19].

DNMT3 enzymes deposit methyl groups regardless of the genomic sequence, thus being responsible for the establishment of DNA methylation patterns after embryo implantation and cell specification. DNMT3 is also responsible for laying methylation in the context of promoter CpG dinucleotides, which are cytosines attached to guanines through a phosphodiester bond. On the other hand, DNMT1 is involved in the maintenance of the genomic methylation profile through its action in newly synthesized DNA. During replication, DNMT1 is recruited to hemi-methylated CpGs at replication forks in order to methylate the daughter strand, keeping the CpG methylome. Therefore, only symmetrical CpG methylation is maintained during replication[18,20].

1.2.3.1. Biological role of DNA methylation – a focus on transcription

DNA methylation is widespread across the mammalian genome. In human somatic cells, is it estimated that around 4% of cytosines are methylated, and it is reported to happen almost exclusively in paired symmetrical CpG sites[20]. However, in the brain tissue, methylation is vastly found in non-CpG contexts. For instance, a postnatal peak in CpA methylation has been reported[18]. Furthermore, non-CpG 5mC is significantly observed in embryonic stem (ES) cells. DNA methylation is a critical regulatory layer of mammalian embryogenesis and germline development. These processes encompass strong DNA methylome reprogramming, which provides the gene plasticity required for differentiation[18,20]. Although DNA methylation profoundly affects embryogenesis, its effects are not restricted to the specific context of development. Indeed, DNA methylation exerts broad effects on gene expression regulation, which are region-dependent, as discussed below.

Approximately 70% of human genes' promoters are composed of CpG islands (CGIs), which are stretches of variable length (average of 1 kbp) rich in CpG dinucleotides. These promoters may be repressed through H3K27 methylation, creating a more fluid and easy-to-revert inactive state. However, strong silencing of CGI promoters happens mainly through DNA methylation, as illustrated by the robust correlation between CGIs methylation and transcription repression[18,21]. Usually, methylation-dependent gene silencing occurs via impairment of transcription factors binding to their DNA motifs. That is why regions of accessible chromatin are frequently devoid of methylation, as supported by evidence associating low-methylated regions with DNase-I-hypersensitive (DHS) sites[22]. Some well described processes that require solid gene silencing through DNA methylation in somatic

tissues are X-chromosome inactivation, genomic imprinting, and repression of germ-line specific genes. Furthermore, the methylation-dependent suppression of transposable elements, particularly retrotransposons, is of utmost importance given the vastness of our genome occupied by these repetitive elements. Intuitively, CGI promoters of actively transcribed genes are usually devoid of DNA methylation[18]. However, exceptions to this repressive effect have also been reported, as exemplified by the methyl-CpG binding protein 2 (MeCP2), a transcription factor that reveals a preference for mCpG-containing sequences[20,23].

Besides blocking the accessibility of transcription factors, DNA methylation can also drive gene silencing by promoting the assembly of heterochromatin. Indeed, both DNMTs and 5mC are able to recruit chromatin remodelling proteins that have an impact on nucleosome positioning and histone modifications, leading to repression of transcription[18]. For instance, the presence of DNA methylation is necessary and sufficient to direct the assembly of a repressive chromatin state, characterized by histone H4 deacetylation and H3K9 methylation, while preventing H3K4 methylation[24]. Furthermore, the Ubiquitin-Like-Containing PHD And RING Finger Domains 1 (UHRF1) protein, important for targeting DNMT1 to replication forks and thus DNA methylation maintenance, relies on the combined binding to hemi-methylated CpGs and methylated H3K9 (H3K9me2/3)[25]. In mice, loss of Uhrf1's H3K9me2/3-binding activity resulted in a moderate reduction of DNA methylation levels across several tissues. Accordingly, *in vitro* experiments demonstrate that H3K9 methylation enhances UHRF1 binding to nucleosomes: UHRF1 binds preferentially to H3K9me2/3-containing nucleosomes[26].

In contrast to active promoters, the bodies of actively transcribed genes are DNA methylation-rich, a pattern highly conserved across eukaryotes. One of the DNMT3 protein domains binds H3K36me3 *in vitro*[27], which is a histone mark that arises as RNA Polymerase II (RNA Pol II) elongates[28]. This evidence suggests a connection between *de novo* DNA methylation and co-transcriptional regulation. In agreement, studies have shown that DNMT3B-dependent gene body methylation in mouse ES cells requires the active transcription mark H3K36me3[20]. Additionally, DNMT3A mediates gene body methylation during post-natal neurogenesis, supporting the role of DNA methylation as a regulatory platform in co-transcriptional regulation[20].

Splicing is a prominent example of such co-transcriptional processes. DNA methylation has been directly implicated in splicing regulation through changes in RNA Pol II kinetics. The recruitment of transcriptional repressors that bind DNA in a methylation-

sensitive manner will favor or impair exon inclusion levels. It is the case of CCCTC-binding factor (CTCF), whose binding to the exon 5 of *CD45* is inhibited by DNA methylation, leading to exon exclusion[29]. The opposite effect has also been observed, for instance through the previously mentioned MeCP2, which is specifically targeted to methylated exons, allowing their recognition and inclusion[30].

Such implications of DNA methylation have been further explored in alternative splicing, which enables the creation of more than one unique mRNA species from a single gene. It is estimated that around 95% of human multi-exon genes undergo alternative splicing[31]. There are several types of alternative splice events contributing to mRNA diversity and proteome enrichment. Some of the most common consist of exon skipping (exclusion of specific exons, referred to as cassette exons), use of alternative splice sites (which affects the boundaries between introns and exons), intron retention, and variations of mRNA untranslated regions (UTRs)[32]. On average, introns are less methylated than exons, and among these, constitutive exons exhibit higher methylation levels than alternative exons, suggesting that DNA methylation promotes inclusion. A study in ES cells lacking DNA methyltransferase activity revealed that DNA methylation affects the splicing of more than 20% of alternative exons[33]. Indeed, DNA methylation seems to fine-tune the splicing of exons whose splicing sequences are not strong enough for recognition by the splicing machinery. In these cases, DNA methylation can influence alternative splicing in both directions (either enhancing or impairing exon inclusion), and this opposite impact is correlated with DNA methylation levels: alternative exons that are negatively affected by methylation (excluded) have significantly higher levels of methylation than positively affected exons (included)[33]. These dichotomous outcomes indicate that DNA methylation does not have a linear effect on splicing, consisting most likely of yet another regulatory layer of the complex splicing mechanism, which depends on several other factors.

1.2.3.2. DNA demethylation by TET enzymes

DNA demethylation is carried out by the ten-eleven translocation (TET) family of proteins. Their name derives from the chromosomes ten-eleven translocation observed in rare cases of acute myeloid and lymphocytic leukaemia, characterized by the fusion of the mixed-lineage leukaemia 1 (*MLL1*) gene in human chromosome 10, with the *TET1* gene on human chromosome 11. TET enzymes are present in all metazoans that have retained cytosine methylation. In mammals, TET family comprises TET1, TET2 and TET3[34]. TETs

are dioxygenases that catalyse the transfer of oxygen atoms from molecular oxygen to organic substrates, and require Fe^{2+} and 2-oxoglutarate as cofactor[34]. TETs were initially reported to oxidize 5mC to 5-hydroxymethylcytosine (5hmC)[35]. Only later TETs were shown to also catalyse the successive oxidization of 5hmC to 5-formylcytosine (5fC) and 5-carboxylcytosine (5caC)[36]. These oxidized cytosine forms can engage in different pathways that lead to their replacement by native cytosine[34], as described below and illustrated in Figure 2.

1.2.3.2.1. Active demethylation

The best-characterized mechanism of active DNA demethylation involves the base excision repair (BER) DNA repair process, which replaces damaged or mismatched bases to avoid mutations or DNA breaks during replication. Briefly, TETs further oxidation products 5fC and 5caC undergo base excision by thymine DNA glycosylase (TDG), creating an abasic site that will be processed to integrate a native cytosine[37,38]. Actually, *in vitro* assays have demonstrated that TDG binds to 5caC:G with higher affinity than to its conventional substrate T:G[37]. Genomic 5fC and 5caC are rapidly removed by TDG, and even in TDG-deficient cells, they are still much less represented than 5mC (around 0,2% of total 5mC), suggesting these modifications are short-lived intermediates of DNA demethylation[34].

Although more controversial, a deamination-dependent DNA demethylation mechanism has also been proposed. In one scenario, 5mC is deaminated by activation-induced cytidine deaminase (AID) and apolipoprotein B mRNA editing enzyme complex (APOBEC), generating a T:G mismatch that is subsequently repaired by TDG/ BER[39]. Alternatively, AID and APOBEC-mediated deamination of 5hmC creates 5-hydroxyuracil (5hmU), which is then removed by single-strand-selective monofunctional uracil DNA glycosylase 1 (SMUG1) or TDG, and reverted to cytosine[40]. However, because AID preferentially acts on single-stranded DNA (ssDNA)[41] and AID and APOBEC enzymes are primarily efficient on unmodified cytosine substrates[42], these deamination-dependent demethylation mechanisms are quite debatable.

Some studies have hypothesized other non-conventional DNA demethylation processes, namely through direct enzymatic decarboxylation of 5caC, or via DNMT-mediated dehydroxymethylation of 5hmC. However, these mechanisms still lack stronger evidence[34].

1.2.3.2.2. Passive demethylation

As previously described, maintenance of methylation patterns during replication requires DNMT1 to methylate the daughter strand at hemi-methylated CpG sites formed at replication forks. However, the recruitment of DNMT1, as well as its enzymatic activity, are severely hampered in the context of hemi-hydroxymethylated DNA in comparison to hemi-methylated. Therefore, TET-mediated hydroxymethylation of mCpG sites critically impairs the DNMT1-dependent maintenance of methylation during replication, leading to a passive loss of 5mC with progressive cell division[34].

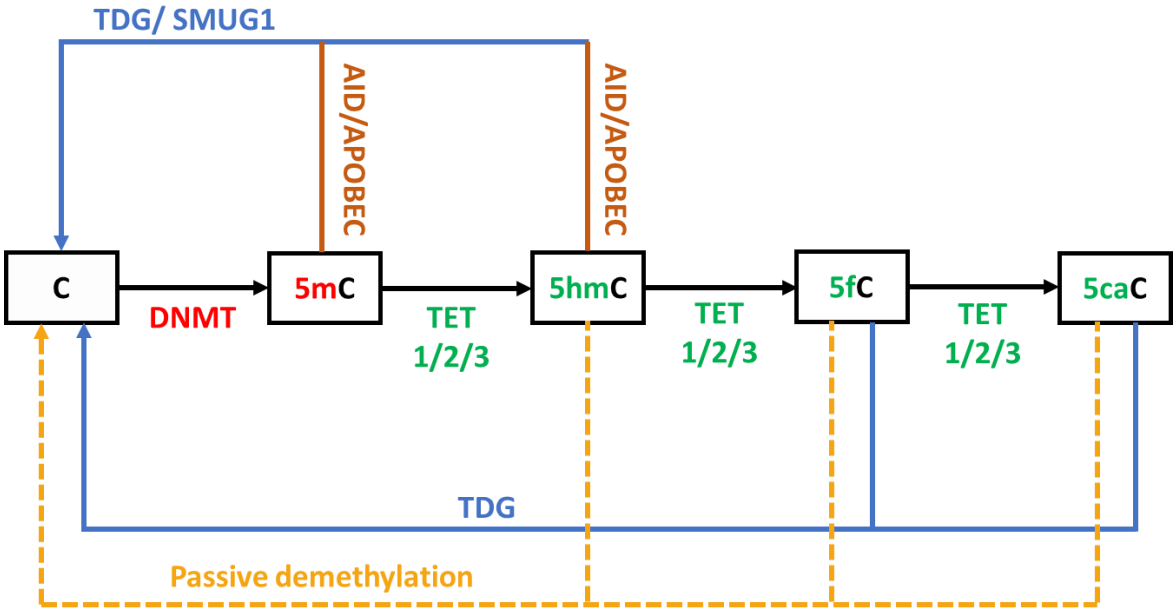


Figure 2: DNA demethylation pathways. After DNMT-mediated conversion of cytosine in 5mC, TET enzymes catalyze the successive oxidation of 5mC in 5hmC, followed by 5fC and finally 5caC. Each of these oxidation products may engage in different pathways that lead to their replacement by native cytosine. The best-described mechanism of active demethylation relies on base excision repair: 5fC and 5caC are subject to TDG, creating abasic sites that are processed to incorporate native cytosine. Although more controversial, a deamination-dependent demethylation process has also been proposed, in which AID/ APOBEC-mediated deamination of 5mC or 5hmC generates mismatched intermediates, which are then processed by TDG or SMUG1. Moreover, all TET-driven oxidation products facilitate passive DNA demethylation[34].

1.2.4. TET enzymes – structure, affinity and redundancy

Towards the amino terminus, TETs contain a DNA-binding CXXC domain with Zn^{2+}-chelating characteristics featuring $CGXCXXC(X)_NC$ amino acid signature sequence, where C refers to cysteine, G to glycine and X represents any amino acid. In metazoans, CXXC domains contain two of these sequences. Regarding the carboxy-terminal, it is composed of a catalytic core region that includes a cysteine-rich insert and a larger double-stranded β-helix (DSBH) domain[34]. *TET1*, *2* and *3* result from a triplication that the *TET* gene suffered in jawed vertebrates. Then, *TET2* underwent a chromosome inversion in which the exon containing the CXXC domain was detached from the catalytic domain coding region, becoming a separate gene that encodes the Inhibition of the Dv1 and Axin complex (IDAX) protein (also known as CXXC4). Although *IDAX* and *TET2* genes are transcribed in opposite directions, the IDAX CXXC domain interacts directly with the catalytic domain of TET2[43].

TETs' DNA-binding CXXC domains occur in many chromatin-associated proteins and they are divided into three subfamilies according to their sequence. TET1, TET3 and IDAX CXXC domains all fall in subfamily 3. Nonetheless, *in vitro* studies indicate that TETs present different DNA binding affinities[34]. *In vivo*, all three TET enzymes are mainly targeted to CGIs in CpG-rich promoters and exons[43–45]. Additionally, the cysteine-rich insert present in the carboxy-terminal region of all TETs has been reported to help in target recognition, as part of a DNA-binding surface[34].

The redundancy of TETs enzymatic activity is still under debate. The differential binding affinities of TET1,2 and 3 CXXC domains, together with the cysteine-rich DNA binding platform shared by the three enzymes, are illustrative of their dual action as orchestrated transcription factors that regulate distinct targets, while simultaneously exerting interchangeable roles. For instance, studies have demonstrated TETs' individual roles in the stepwise oxidation reactions from 5mC to 5caC that take place during DNA methylation reprogramming in the mouse zygote[46]. Conversely, TETs partial redundancy as tumour suppressors has been reported in the mouse haematopoietic system, specifically in preventing oncogenic events that drive myeloid malignancies[47]. Therefore, evidence suggests that TETs may play both interchangeable and non-redundant roles, depending on the cellular event and physiological context.

1.2.5. Biological role of DNA hydroxymethylation

The role of 5hmC in physiological processes is still vastly unknown, but several studies acknowledge 5hmC as a new mark in the epigenetic landscape that directly influences genome structure, chromatin organization and gene transcription regulation, rather than simply being a transient intermediate of the demethylation process[48].

1.2.5.1. 5hmC along the gene – a focus on transcription

Due to the low frequency of 5hmC, high sensitivity methods are required to accurately map this modification across the genome. Bisulfite sequencing, the standard technique used to draw methylation patterns, does not allow to distinguish 5mC and 5hmC, since both marks are resistant to deamination by sodium bisulphite[34]. Therefore, several methods have been employed to map 5hmC, which generate sometimes contradictory results due to their inherent limitations. For instance, a long DNA stretch containing conserved but sparse and dispersed 5hmC marks might be recognized as 5hmC-rich through antibody-based immunoprecipitation protocols, while single-base resolution methods consider it to be mostly devoid of 5hmC[34,49].

Nonetheless, there is consensus on the distinctive 5hmC distribution features across some important regulatory regions. 5hmC mapping throughout a panel of different human tissues revealed that it is deficient at transcription start sites (TSSs) but enriched at promoters and exons[49,50]. In contrast to 5mC, which is usually devoid from promoter regions, 5hmC enrichment at gene promoters associates with an active transcription state. For instance, studies have shown that ES cells, neurons and neural progenitor cells present high promoter 5hmC levels[34].

Moreover, 5hmC enrichment in gene bodies is widely observed across cell types, where it is usually positively correlated with gene expression. This has been reported in ES cells, liver and brain tissues, and even in some cancer cells. In neural progenitors and neurons, activation of neuronal function-related genes leads to 5hmC accumulation preferentially in gene bodies, while poorly expressed genes show intragenic depletion of 5hmC[34,51].

Although enhancers usually display relatively low CpG density and DNA methylation, 5hmC enrichment is observed in these regions, indicating that they are subject to strong TET activity. In agreement, evidence shows an accumulation of 5fC and 5caC at

enhancers in TDG-depleted ES cells[34]. 5hmC at enhancers is predominantly located next to transcription factor-binding sites, but absent from the site of binding, suggesting that 5hmC favors transcription factor accessibility to DNA. Opposite to neurons, ES cells display higher 5hmC content in enhancers than in gene bodies. Strikingly, enhancers' hydroxymethylation peaks in differentiating ES cells immediately after the onset of differentiation, in parallel with the gain of active enhancer histone marks, such as the acetylation of H3K27[34,51], suggesting a dedicated role of 5hmC in the differentiation process.

1.2.5.2. 5hmC in ES cells and neurons

There is a high discrepancy in 5hmC abundance and genomic distribution among cell types, which is likely the reflection of differences in 5mC prevalence and cell-specific regulation of TETs enzymatic activity. In average numbers, 5hmC percentage across the genome is approximately 0,1-0,7% (about 10-100 fold decrease from what is typically observed for 5mC)[50]. From all tissues, it is the brain that displays the highest 5hmC levels, reflecting the accumulation of 5hmC in the post-mitotic cells of the nervous system, where 5hmC is not erased in a passive, replication-dependent manner. For instance, in Purkinje neurons, 5hmC reaches 40% of the total 5mC levels[34]. In some immune cell populations, 5hmC represents about 1% of total 5mC, and in ES cells it represents 5-10%[34]. The abundance of 5hmC in the brain's complex neural circuitry (and its postnatal increase in different brain regions), as well as in ES cells that require extensive gene plasticity, are suggestive of a key role in the methylation-dependent regulatory network[51,52].

Indeed, the DNA methylome is dynamically reprogrammed during early embryonic and germ cell development[53]. TET proteins are strongly implicated in the development stages during which mass demethylation takes place, namely in the early zygote, immediately after fertilisation, and in primordial germ cells of the developing embryo (the precursors of mature germ cells)[34]. In agreement, ES cells display high *TET* expression levels, as reflected by the enrichment of 5hmC at gene promoters and CGIs, where it correlates with increased transcriptional levels, while 5mC is underrepresented in these regions[53]. TET enzymatic activity is critical to preserve ES cell maintenance and self-renewal capacities, as well as for inner cell mass specification[54]. Abrogation of TETs disrupts the methylation landscape, leading to increased methylation and silencing, which directly affects ES cell-specific genes, such as pluripotency and lineage commitment genes[53,54]. For instance, studies reported downregulation of pluripotency-related genes upon

depletion of *Tet1* and *Tet2* in mouse ES cells, causing a bias towards extraembryonic lineage differentiation[53]. Similarly, *Tet1*-depleted pre-implantation embryos have a higher propensity for trophectoderm commitment[54]. Hence, genomic (hydroxy)methylation tightly controls the balance between pluripotency and lineage commitment[53].

Moreover, 5hmC has been described as a stable and functional epigenetic marker in the neurogenesis process. 5hmC increases during neuronal differentiation, and it is highly enriched in active genomic regions, where its incidence positively correlates with gene expression. Also, 5hmC enrichment is not associated with significant DNA demethylation, positioning 5hmC as a stable epigenetic mark in the genome[51]. Remarkably, during neurogenesis, the increase in 5hmC negatively correlates with repressive histone marks, such as H3K27me3, while it is accompanied by active histone modifications, including H3K4me1 and H3K9ac. These data suggest a role for 5hmC as a transcription activator that crosstalks with different epigenetic mechanisms in the fine-tuned gene regulation that occurs during neurogenesis[52,55].

1.3. 5mC and 5hmC influence chromatin structure and thermodynamics

In contrast to 5mC repressive features, the above-mentioned evidences establish 5hmC as a long-term stable and independent epigenetic mark, intrinsically correlated with gene expression activation[55]. Thus, as with other epigenetic modifications, 5mC and 5hmC regulate gene expression by altering the structure and functionality of chromatin, therefore allowing coordinated access of transcription factors, polymerases and other protein complexes (Figure 3)[56].

1.3.1. Influence on nucleosome dynamics

As previously described, the eukaryotic genome's repetitive unit is the nucleosome, composed of two of each H2A, H2B, H3 and H4 histones, which form an octamer that is wrapped by 147 bp of DNA. $(H3\text{-}H4)_2$ tetramer compose the nucleosome's core, which is flanked by H2A-H2B dimers[57]. *In vitro* biochemical studies have already shed light on how 5mC and 5hmC impact nucleosome assembly, compactness and stability. Results demonstrate that hydroxymethylation facilitates nucleosome formation, while methylation significantly reduces the binding affinity of DNA to histone octamers. However, once

assembled, hydroxymethylation may lead to more dynamic nucleosome conformations, depending on the DNA sequence context. These hydroxymethylation-induced effects are unlikely to change the static conformation of chromatin *in vivo*, but moderate alterations occur in dynamic features such as the DNA-end breathing motion, a primary dynamic feature of the protein–DNA complex, thought to control DNA accessibility[50]. Therefore, hydroxymethylation may facilitate access of transcription machinery to DNA and thus favor transcription initiation[50].

Moreover, during transcription, nucleosome transient unfolding is required to accommodate the passage of RNA Pol II. However, evidence shows that the displacement of the total histone octamer from the DNA is not always necessary. Actually, a mechanism has been proposed in which removal of only the H2A-H2B dimer flanking the tetramer is sufficient for transcription machinery to go through[57,58]. In support of this, higher exchange rates of H2A-H2B histones compared to H3 and H4 have been reported, as well as the close correlation observed between the presence of reactive nucleosomes (as those formed upon loss of H2A-H2B dimer) and ongoing transcription[57]. Strikingly, 5hmC significantly weakens the interaction of DNA with the H2A-H2B dimers, thus creating an open and active chromatin state that reinforces its function in active transcription[50].

1.3.2. Thermodynamic effect on DNA melting temperature

In addition to the impact of DNA (hydroxy)methylation on nucleosome dynamics, studies already provided insights on how these marks influence thermodynamic features of double-stranded DNA (dsDNA) molecules[56,59]. Evidences have demonstrated that 5mC and 5hmC exert opposite forces on dsDNA thermodynamic stability by affecting DNA melting temperature. Studies benefiting from high-resolution melting analysis determined DNA melting curves of an 897 bp DNA fragment, with even distributions of G, A, T and C, and in which all cytosines are either native, methylated or hydroxymethylated. Taking as reference the temperature necessary to reach 50% denaturation, 5mC caused an increase in DNA melting temperature of around 6ºC compared with the unmodified fragment, while 5hmC led to a decrease in melting temperature of around 2ºC. Moreover, these effects were also demonstrated at the single base modification level, using a 52 bp fragment with a single C-modified nucleotide. Although less striking, the same trend was observed: 5mC increases while 5hmC decreases dsDNA melting temperature[59]. Identical results were reported in another study, describing increasing annealing temperatures from 5hmC to C to 5mC

fragments[56]. These data positions 5hmC-modified dsDNA as less energetically stable and more easily perturbed, as compared to the 5mC- or C-containing DNA molecules.

1.3.3. Impact on C:G intra-base-pair fluctuations

Regarding local DNA structure and geometry, molecular dynamics simulations were performed to determine how cytosine modifications affect local C:G pairing. Measurements of the amplitude of intra-base-pair fluctuation criteria (such as shear, stretch, stagger and buckle) showed higher fluctuation amplitudes in 5hmC-modified fragments, followed by C and finally 5mC[56]. Intra-base-pair fluctuations depend on the interplay between steric effects (dictated by the modification size) and the modification polarity. Modification's increasing size commonly leads to higher base-pair rigidity, while increasing polarity renders it more flexible by promoting solvent-mediated fluctuations. According to this, the hydrophilic hydroxymethyl group destabilizes G-5hmC pairing by alleviating steric constraints through an increase in molecular polarity, while the methyl group stabilizes G-5mC pairing due to the combined steric effect and hydrophobicity[56].

1.3.4. Effect on dsDNA global structure and rigidity

Besides the local effect on DNA fluctuations, cytosine modifications also have a significant impact on global DNA rigidity, as determined through atomic force microscopy (AFM) experiments using the end-to-end distance of DNA fragments immobilized on a surface as a read-out for global rigidity (the longer the distance, the higher the rigidity)[56]. Results revealed that the mean end-to-end distance was significantly shorter for 5hmC-DNA, followed by C-DNA and then for 5mC-DNA molecules, which displayed the longer measurements. This evidence supports the increased flexibility induced by hydroxymethylation, in contrast with the higher rigidity caused by methylation, not only at the level of C:G paring but also at the global dsDNA molecule context[56].

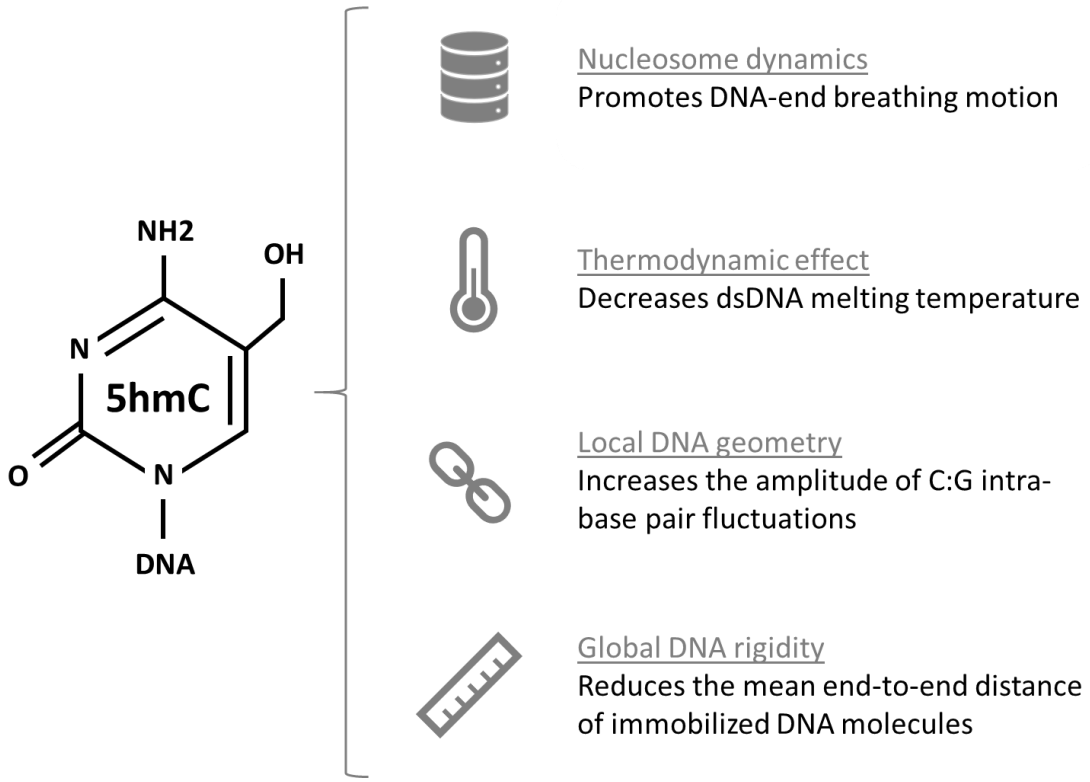


Figure 3: 5hmC creates a more open and active chromatin state by altering its structure and functionality. 5hmC-containing DNA yields nucleosomes with facilitated DNA-end breathing motion, a dynamic feature of the protein–DNA complexes thought to control DNA accessibility, as well as weaker interactions between DNA and nucleosomal H2A-H2B dimers, facilitating RNA Pol II elongation[50]. Also, 5hmC diminishes the thermodynamic stability of the DNA duplex, causing a decrease in dsDNA melting temperature when compared to native C, in contrast to 5mC that increases DNA melting temperature[56,59]. Regarding local DNA structure and geometry, molecular dynamics simulations revealed that the highest amplitude of C:G DNA base-pair fluctuations is observed in the presence of 5hmC, whereas 5mC yielded the lower amplitude values[56]. This increased flexibility induced by hydroxymethylation was also reported at the global dsDNA molecule level, using the end-to-end distance of immobilized DNA fragments as a read-out for global rigidity, which revealed that the mean end-to-end distance was significantly shorter in the presence of 5hmC[56].

1.3.5. 5mC and 5hmC as DNA molecular switches that impact genome integrity

The above-mentioned findings emphasize the inverse roles of 5mC and 5hmC as repressors or activators of transcription, respectively. On the one hand, 5mC increases DNA melting temperature[56,59], stabilizes DNA base-pairing[56] and leads to an overall higher DNA rigidity[56], impairing DNA unwinding[50], which is required for transcription initiation and elongation. On the other hand, 5hmC reduces DNA melting temperature[56,59], destabilizes DNA duplexes[56] and establishes a more relaxed chromatin state[56], favoring transcription factor binding and RNA Pol II elongation[50]. These direct effects of 5mC and 5hmC on DNA

thermodynamic stability and chromatin structure characterize them as DNA intrinsic molecular switches, which affect not only transcription regulation, but all chromatin-based events, such as DNA damage and repair[60].

Indeed, the open chromatin conformation instructed by 5hmC, which contrasts with the 5mC-associated chromatin compaction, suggests that this epigenetic mark might exert a regulatory role during DDR, for instance through the establishment and maintenance of a local chromatin landscape that allows access and/ or recruitment of repair machinery[60]. In agreement, data shows 5hmC accumulation at DNA damage sites induced by aphidicolin or microirradiation in HeLa cells, where it colocalizes with characteristic DNA damage markers as γH2AX, and helps prevent chromosome segregation defects in response to replication stress [60]. Thus, 5hmC is directly linked to genome integrity.

A major source of co-transcriptional DNA damage are R-loops, triple-stranded structures formed during transcription when the nascent RNA molecule hybridizes with the template DNA, forming a DNA:RNA hybrid and displacing the coding ssDNA. Non-scheduled or persistent R-loops are drivers of DNA damage, namely DNA double-strand breaks (DSBs). Thus, these structures require a fine-tuned regulation in order to preserve genome integrity[61].

1.4. R-loops: what are they? How do they form?

DNA:RNA hybrids are for long known to form during DNA replication (the 11 bp Okazaki fragments) and during transcription (the 8 bp hybrid within the RNA Pol II active site). However, longer tracts of co-transcriptional DNA:RNA hybrids also occur when the nascent RNA hybridizes with the template DNA, creating R-loops[62]. These triple-stranded structures are found in the genome of a variety of organisms, such as bacteria, yeast, or mammals, and even in organelles, specifically mitochondria[63,64].

Since the formation of a short and transient DNA:RNA hybrid takes place within the active core of elongating RNA Pol II, R-loops formation was initially proposed to happen through the "extended RNA:DNA hybrid" model, in which the R-loop would be an extension of this 8 bp DNA:RNA segment within the transcription bubble[65,66]. However, recent studies shed light on the structure of transcribing RNA Pol II and the complex it forms when interacting with nucleic acids. Interestingly, high-resolution cryo-electron microscopy showed that the nascent RNA and the template DNA exit RNA Pol II active core through physically separate channels, demonstrating that R-loops cannot be simply formed as an

extension of the 8 bp hybrid[67,68]. Nowadays, the most accepted model for R-loop formation is the "thread back model", which proposes threading back of nascent RNA with template DNA before the two strands of DNA reanneal. Although physically separated, nascent RNA is still close to template DNA, which allows the invasion of the DNA duplex, creating an R-loop. R-loops adopt an intermediate conformation between the B-form of dsDNA and the A-form of double-stranded RNA (dsRNA), which grants them higher thermodynamic stability than dsDNA[63,65].

1.4.1. R-loop homeostasis

Some intrinsic features of the transcribed DNA are known to favor R-loop formation[69]. However, given their potentially harmful effect on genome stability, cells have acquired a variety of mechanisms to maintain R-loop homeostasis (Figure 4). These mechanisms can be divided into two groups: those that prevent R-loop formation and those that resolve already formed R-loops[61].

1.4.1.1. R-loop favoring genomic features

R-loop formation is directly influenced by genomic GC content, DNA supercoiling and DNA cleavage. Regions with strong G clustering favor R-loop formation[69]. Specifically, a strand asymmetric distribution of Gs and Cs (termed GC skew) with enrichment of Gs over Cs in the coding strand (positive GC skew) promotes R-loops[70], which are further stabilized through the establishment of G-quadruplex structures on the displaced G-rich ssDNA[65]. G-quadruplexes are tertiary nucleic acid structures formed in G-rich regions. Their basic unit is the G-quartet, a planar array of guanines stabilized by Hoogsteen base-pairing, which can stack on top of each other to create G-quadruplexes[71]. Regarding DNA supercoiling, the negative supercoiling accumulating upstream of a transcribing RNA Pol II creates a local DNA unwinding that provides a window of opportunity for nascent RNA to invade the DNA duplex and hybridize with the template DNA strand[63,65]. Additionally, there is evidence on how a DNA nick can serve as an R-loop initiation site by resecting the beginning of the transcript[72].

1.4.1.2. Mechanisms to prevent R-loop formation

Since most RNA processing events occur co-transcriptionally (as capping, splicing, editing and export)[73], R-loop formation can be prevented through the action of RNA-binding proteins. The ligation of such factors to nascent RNA creates a ribonucleoprotein (RNP) complex that physically inhibits RNA invasion of the DNA duplex. This is the case of the Transcription and Export (THO/TREX) complex in yeast, which couples transcription and pre-mRNA maturation with mRNA export[74,75]. Furthermore, the quick ligation of early splice-site recognition factors to the emerging RNA also hinders DNA:RNA hybridization[76]. Indeed, serine/arginine splicing factor 1 (SRSF1)-depleted cells revealed a hyper-mutagenic phenotype caused by an accumulation of R-loops[77].

Since R-loops are highly favored by the negative DNA supercoiling created behind elongating RNA Pol II, another mechanism to withhold R-loop formation consists in alleviating this topological tension[63]. Topoisomerases, which can relieve DNA torsional tension[78], prevent co-transcriptional R-loop formation by relaxing negatively supercoiled DNA[79]. Indeed, evidences show that topoisomerase I protects genome integrity in mammalian cells by suppressing the accumulation of co-transcriptional R-loops, as well as preventing replication-transcription conflicts[80].

1.4.1.3. Mechanisms to resolve R-loops

To resolve already formed R-loops, cells rely on the enzymatic activity of nucleases or helicases, which will either digest or unwind the hybrid, respectively[66]. The best characterized R-loop-resolving nuclease is ribonuclease H (RNase H), which specifically digests the RNA moiety in DNA:RNA hybrids. Two types of RNase H enzymes, with different physiological functions, have been described in eukaryotes and bacteria: RNase H1 and RNase H2[62]. In mammals, the RNase H2 enzyme complex is composed of three separate proteins, in contrast to the prokaryotic RNase H2 that functions as a single protein[81]. Regarding the cleavage pattern, RNase H1 requires a substrate with at least four ribonucleotides for cleavage to occur[82], while RNase H2 removes single mis-incorporated ribonucleotides[83]. RNase H1 locates in the nucleus, digesting co-transcriptional R-loops. Additionally, it is also found in mitochondria, where it has been implicated in mitochondrial DNA (mtDNA) replication during embryogenesis (besides resolving transcription-generated R-loops as well)[65,82]. In turn, RNase H2 acts mainly in removing ribonucleotides that are

frequently mis-incorporated in DNA during replication, taking part in the ribonucleotide excision repair (RER) pathway[84].

Therefore, RNase H enzymes are crucial to keep DNA:RNA hybrids in check, as reinforced by the significant increase in chromosome instability that arises in RNase H1 and H2 deficient cells[83]. Moreover, the vital role of RNase H1 in mtDNA replication is demonstrated by the developmental arrest and embryonic lethality observed in *Rnaseh1* null mice, due to a severe decrease in mtDNA content[85]. The fine-balance between R-loops and RNase H is further confirmed in *S. pombe*, where deletion of RNase H stabilizes R-loops around DSBs, impairing DDR pathways, while overexpression of RNase H1 disrupts the hybrids and cause severe loss of repetitive regions around DSBs[86]. Indeed, overexpression of nuclear RNase H1 is the most used and efficient experimental method to diminish cellular R-loop levels[65].

Concerning R-loop resolution via hybrid unwinding, prokaryotic and eukaryotic cells employ several RNA:DNA helicases that are able to unwind the RNA from R-loops, allowing for the reannealing of the two DNA strands. Examples are the bacterial RecG DNA helicase[87], the yeast Pif1p DNA helicase[88], or the human aquarius (AQR) RNA helicase and Senataxin (SETX) RNA/DNA helicase (and its yeast homolog Sen1)[65,66]. Moreover, the human RNA/DNA helicase DEAH box protein 9 (DHX9) was shown not only to unwind RNA:DNA hybrids *in vitro*, but also to resolve DNA-based G-quadruplex structures (an indirect way of regulating R-loops)[89]. As expected, suppression of these helicase enzymes leads to R-loop-dependent DNA damage[66,87].

Given the multitude of R-loop preventing and dissolution factors, they are likely temporally and spatially coordinated to regulate and constrain R-loop homeostasis throughout cell cycle stages and during specific cellular events[62].

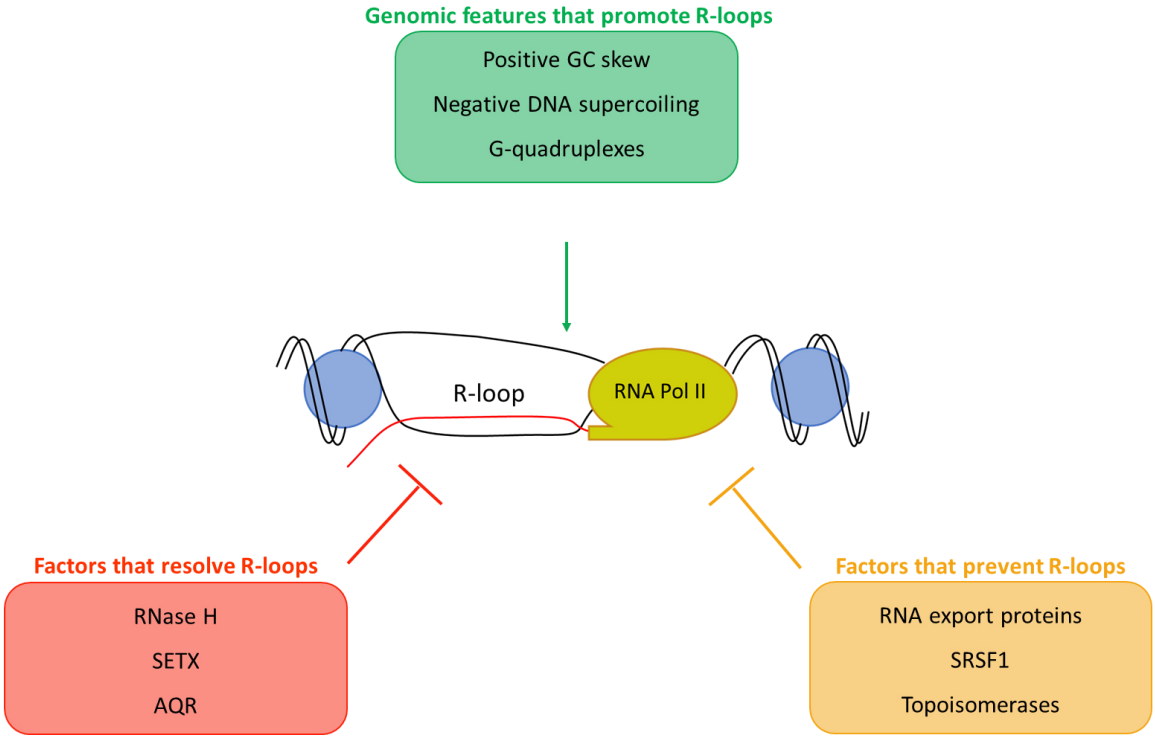


Figure 4: R-loop homeostasis: formation and resolution. Intrinsic features of the transcribed DNA favor the formation of R-loops, such as an asymmetrical distribution of guanine and cytosine nucleotides in the DNA duplex, with an excess of Cs in the template strand (positive G:C skew)[70], which allow the establishment of G-quadruplex structures in the coding strand, further stabilizing R-loops[65]. The negative DNA supercoiling accumulating upstream of a transcribing RNA Pol II also facilitate R-loop formation by creating a local DNA unwinding that promotes RNA invasion of the DNA helix[63]. Cells must therefore employ mechanisms to restrict non-scheduled or persistent R-loop formation, which can be divided in factors that resolve R-loops and factors that prevent their assembly. R-loops can be resolved for instance by the AQR RNA helicase or the SETX RNA/DNA helicase, that remove R-loops by unwinding the hybrids, or by RNase H ribonucleases, which digest the RNA moiety in DNA:RNA hybrids[61]. Regarding R-loop-preventing factors, examples are RNA-binding proteins related to mRNA maturation and export, splicing factors that retain the transcript (such as SRSF1), or topoisomerases, which alleviate topological tensions of the transcribed DNA[61].

1.4.2. Biological role of R-loops

The role of R-loops as intermediates in some specific cellular processes has soon been acknowledged, such as the cases of *E.coli* plasmid replication, mtDNA replication and immunoglobulin class switching[61]. Other than that, R-loops used to be considered rare events, transcription by-products without specific cellular impact. However, during the last two decades, the role of R-loops' has been described in important physiological processes. Examples are gene expression, DNA repair, telomere regulation or even chromatin structure and remodelling[63,90]. Contributing to this was the genome-wide mapping of R-loop occupancy through DNA:RNA hybrid immunoprecipitation, followed by high-throughput

sequencing[91]. R-loops were characterized as structures of variable length, ranging from a few hundred base-pairs up to over 1 kbp. Additionally, R-loops were found to occur at unanticipated high frequency and over conserved loci across thousands of mammalian genes, pointing towards a role in transcription[69]. Nevertheless, along with this new perception, R-loops were also positioned as threats to genome integrity [61].

1.4.2.1. R-loops in transcription

R-loop mapping revealed that they preferably occur in specific gene regions, namely CGI promoters and around TSSs, as well as near the end of RNA Pol II transcribed genes, surrounding the polyadenylation site (PAS). In fact, R-loops have been demonstrated to impact the dynamics of both transcription initiation and termination[65,69].

1.4.2.1.1. R-loops in transcription initiation

As previously mentioned, approximately 70% of human genes promoters are composed of CGIs, which are very rarely methylated[21]. Interestingly, sequence analysis revealed that GC skew around TSSs is a prominent feature of CGI promoters, from which 75% displayed positive GC skew, an R-loop favoring setting[91]. Therefore, CGI promoters possess a widespread tendency to form R-loops upon transcription initiation, which protects DNA from *de novo* methylation, likely by impairing DNMT enzymes binding. DNA methylation is much more frequent across gene bodies, whose CGIs usually don't display strong GC skew. These evidences highlight a functional role of R-loops in transcription initiation, namely in maintaining the unmethylated state of CGI promoters[70,91].

1.4.2.1.2. R-loops in transcription termination

Terminal R-loops are predominantly found in G-rich termination regions downstream of the PAS[65]. Extensive R-loop formation in transcripts that undergo PAS-dependent cleavage and polyadenylation implicates them in the transcription termination process of this class of human genes[69].

Two models have been widely accepted for RNA Pol II transcription termination: the "torpedo" model and the "allosteric" model. The "torpedo" model states that RNA Pol II slows down over the termination region due to the recruitment of 3′-end cleavage and polyadenylation (CPA) complex to RNA Pol II, upon poly(A) signal transcription. Then,

mRNA is eventually released from chromatin, and RNA Pol II continues transcribing, originating short-lived RNAs. These RNAs are degraded from its 5' end by 5'-3' exoribonuclease 2 (XRN2), which will ultimately displace RNA Pol II from the template strand in a torpedo-like effect[92,93]. The "allosteric" model proposes that transcription termination and mRNA release occur through conformational changes in RNA Pol II upon transcription of the poly(A) signal. According to this model, RNA Pol II does not transcribe further because it is irreversibly inactivated due to structural changes in its catalytic domain[94,95]. Likely, both termination models are actually in place and eventually co-exist[93].

Terminal R-loops contribute to slow down and pause RNA Pol II downstream of the PAS. In this context, RNA Pol II elongation causes the upstream accumulation of negative supercoiling, as well as the creation of a nucleosome depleted region due to transient nucleosome displacement. These factors favor RNA invasion of the DNA duplex, particularly over G-rich sequences capable of forming G-quadruplexes[93]. However, hybrids must be resolved to release the nascent RNA and thus allow its XRN2-mediated degradation, leading to RNA Pol II displacement as suggested by the "torpedo model". R-loops are resolved by the RNA/DNA helicase SETX[96]. Sen1, the yeast homolog of human SETX, has also been implicated in transcription termination[97]. Therefore, R-loops are important for efficient transcription termination, but they require a fine-tuned regulation. If the hybrids are not properly resolved, transcription termination is impaired and R-loop accumulation leads to DNA damage, as observed in Sen1 and SETX-knockdown experiments[65,96,98].

1.4.2.1.3. R-loops in gene bodies

Regarding gene bodies, R-loop formation is attenuated by the presence of introns[99]. Several factors explain this observation: intron removal from pre-mRNA eliminates the complementarity required for hybridization; secondary structures of the intronic transcript obstruct RNA invasion[100]; spliceosome assembly detains the transcript, preventing hybridization[99]. This is further strengthened by genome-wide analysis revealing that human intron-containing genes have lower R-loop levels and so are best protected against R-loop-mediated DNA damage[99]. However, although less extensively, R-loops can still occur in gene bodies. In this case, intragenic R-loops are signalled by RNA Pol II stalling, initiating a cascade that enables the recruitment of DEAD-box helicase 23 (DDX23), an RNA helicase that resolves the R-loop and thus restores transcription[101].

1.4.2.2. Interplay between R-loops and epigenetics

The interplay between RNA and the epigenome is an emerging topic in the field of gene expression regulation. As discussed below, evidences show the role of RNA in the establishment of chromatin domains that regulate gene expression through epigenetic mechanisms[102]. Interestingly, these effects can occur in an R-loop-dependent fashion, as summarized in Figure 5.

1.4.2.2.1. Role of R-loops in heterochromatin assembly

The first association between R-loops and heterochromatin formation was observed in *S. pombe*, where heterochromatic ncRNAs associate with centromeric chromatin to form DNA:RNA hybrids. This leads to the establishment of local heterochromatin, in a process mediated by the RNA-induced transcriptional silencing (RITS) complex. Removal of R-loops causes loss of centromeric heterochromatin, highlighting a causal link between R-loops and RNA interference (RNAi)-directed heterochromatin assembly[103].

Later, a similar mechanism was demonstrated to occur in the termination region of the human *β-actin* (*ACTB*) gene, which has a G-rich pause site at its 3' end. R-loop formation in this region leads to antisense transcripts that hybridize with the nascent RNA to form dsRNA. dsRNA recruits RNAi-dependent gene silencing machinery, such as the endoribonuclease DICER and the Argonaute proteins. Additionally, the G9a HKMT is also recruited, depositing H3K9me2 repressive marks over the *ACTB* termination region. This creates a binding platform for heterochromatin protein 1 (HP1), which maintains the heterochromatic setting. Therefore, this cascade of events reinforces RNA Pol II pausing during efficient termination. Remarkably, enzymatic depletion of R-loops reduces the occurrence of antisense RNAs and the occupancy of DICER, G9a and HP1 in the termination region[104].

Furthermore, experiments in *S. cerevisiae*, *C. elegans* and human cells also implicate R-loops in chromatin compaction via histone H3S10 phosphorylation (H3S10P), a mark of chromatin condensation crucial for genome stability. H3S10P levels increase in cells depleted of the mRNA processing THO Complex 1 (THOC1) or SETX helicase, hence with exacerbated R-loops, and its spatial distribution is tightly linked to hybrid accumulation. R-loops removal suppresses H3S10P, further strengthening the role of these hybrids in

chromatin structure remodelling, in a mechanism thought to be conserved in all eukaryotes, including yeast, nematodes and humans[105].

1.4.2.2.2. R-loops and promoter-proximal chromatin

As previously mentioned, R-loops are enriched around TSSs due to the presence of promoter CGI that highly favor their formation[91]. Those R-loops often occur in the context of promoter-proximal pausing of RNA Pol II, a key regulatory step in early transcription widely observed in metazoans and common to the majority of active genes. It consists in the pausing of transcriptionally engaged RNA Pol II through the association of pause-inducing factors, around 30–60 nucleotides downstream the TSS. Productive elongation requires polymerase release from the pausing site[106,107]. RNA Pol II pausing, together with the negative supercoiling and sequence characteristics of the DNA region, facilitate hybrid formation, further stabilizing paused RNA Pol II through an anchoring effect. Indeed, for most CGI promoters, RNA Pol II pausing positively correlates, both in position and in intensity, with high GC skew, and genes with promoter-proximal paused RNA Pol II show R-loop enrichment over the CGI domain. An example of this is observed upon BRCA2 inactivation, which causes RNA Pol II accumulation at promoter-proximal sites, as well as unscheduled R-loop occurrence and the resulting DNA damage[108]. Productive transcription requires hybrid resolution, or otherwise efficient transcription elongation is impaired[109].

Additionally, R-loops have been directly implicated in chromatin remodelling of promoter-proximal regions in mouse ES cells, affecting gene expression and differentiation. R-loops affect the binding of two chromatin regulatory complexes: the H3K27 methyltransferase polycomb repressive complex 2 (Prc2), important for gene silencing during development, and the Tip60–p400 histone acetyltransferase complex, targeted to nascent transcripts. Genes with promoter-proximal R-loops have increased binding of Tip60–p400, but lower Prc2 levels. Interestingly, R-loop depletion leads to unbalanced Tip60–p400 and Prc2 recruitment genome-wide, and overall impairment of ES cells differentiation[110]. Altogether, these evidences demonstrate the role of R-loops in tailoring promoter-proximal chromatin modifications with a direct impact on the recruitment of key pluripotency regulators[110].

1.4.2.2.3. R-loops and DNA methylation

As previously mentioned, R-loops affect DNA methylation patterning by acting as a protective platform against *de novo* DNA methylation in human CGI promoters with high GC skew[91]. Indeed, a strong correlation between positive GC skewness and an unmethylated epigenetic state becomes evident when comparing promoter CGIs, which display high GC skew and are often unmethylated, with gene body CGIs, presenting poor GC skew and often methylated. Thus, R-loop-forming potential, estimated based on GC skewness, is predictive of the (un)methylated state of CGIs, establishing a direct link between R-loops and the methylome[70].

Recent studies have further strengthened the above-mentioned interplay. For instance, amyotrophic lateral sclerosis 4 (ALS4) patients carry a SETX mutation that enhances R-loop depletion. In patients with this motor neuron disease, overall R-loops are reduced, causing significant changes in gene expression. Specifically, a reduction in BMP and activin membrane bound inhibitor (*BAMBI*) expression, as well as in its promoter R-loops, was observed in fibroblasts from ALS4 patients. BAMBI is a pseudoreceptor that negatively modulates the transforming growth factor β (TGF-β) pathway by preventing the formation of functional receptors. Fibroblasts from ALS4 patients don't express DNMT3 proteins, but DNMT1 is present. These cells show increased DNMT1 binding and methylation levels at *BAMBI* promoter, causing transcription repression. DNMT1-knockdown restores *BAMBI* expression, demonstrating that *BAMBI* silencing is DNMT1-dependent. Strikingly, enzymatic depletion of R-loops further increases promoter methylation, suggesting that the presence of a DNA:RNA hybrid deters methylation by DNMT1. A genome-wide analysis of ALS4 patients revealed that promoter R-loop accumulation is a common protective feature of several genes against methylation silencing and that gene expression changes caused by the disease are partially explained by the disruption of this balance[111].

Another example of the impact of R-loops in gene methylation is found in the vimentin (*VIM*) locus, which in colon cells undergoes both sense and antisense transcription. Interestingly, the *VIM* antisense transcript is involved in the formation of an R-loop in the vicinity of *VIM* TSS, which is required to maintain an open chromatin state that favors transcription factor binding. In agreement, antisense knockdown prevents R-loop formation, causing *VIM* promoter CGI hypermethylation and *VIM* downregulation. Direct R-loop depletion is also sufficient to repress active chromatin and diminish transcription factor

binding. For instance, in colon cancer, *VIM* expression is deregulated due to promoter hypermethylation which silences transcription in both directions. The establishment of an R-loop-driven regulatory platform for methylation and chromatin accessibility in the *VIM* promoter might be extrapolated to other CGI promoters with divergent sense and antisense transcription[112].

Besides deterring DNMTs activity, R-loops are also implicated in active DNA demethylation. The lncRNA TCF21 antisense RNA inducing promoter demethylation (*TARID*) is transcribed in the antisense strand of the tumour suppressor transcription factor 21 (*TCF21*), forming an R-loop that maintains the *TCF21* CGI promoter in an unmethylated and active state. In this case, the R-loop recruits the RNA:DNA-binding protein growth arrest and DNA damage inducible alpha (GADD45A), which in turn tethers TET1 enzyme. Indeed, R-loop formation and GADD45A recruitment coincided with TET1 occupancy and 5hmC presence at the *TCF21* promoter. R-loop depletion causes promoter hypermethylation, *TCF21* and *TARID* downregulation, and reduces GADD45A binding and 5hmC levels. Regarding the genome-wide effect of R-loops in active DNA demethylation, R-loop immunoprecipitated fragments show enrichment of demethylation intermediates, suggesting that these hybrids may mark sites of demethylation machinery activity. However, GADD45A-directed recruitment of TET1 to R-loops is likely a limited effect, as only 4% of TET1 peaks were decreased upon R-loop depletion in mouse ES cells, from which the vast majority localizes near TSSs in CGI regions. Thus, R-loop guidance of TET1 activity occurs mainly at the promoter context[113].

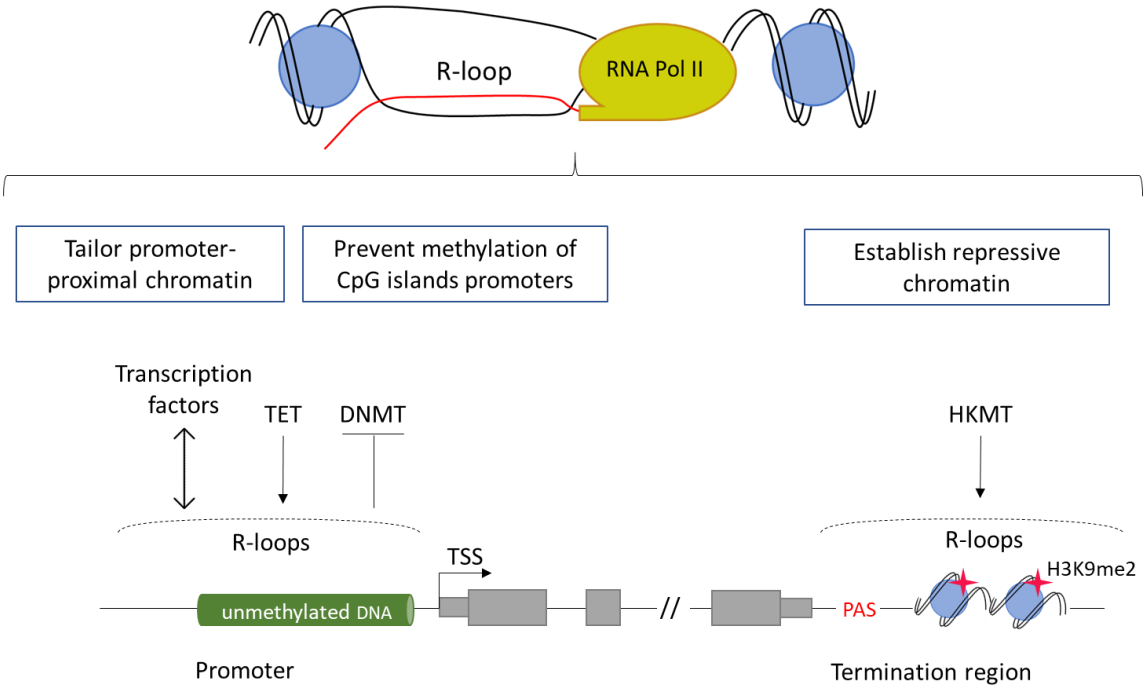


Figure 5: R-loops influence epigenetic modifications that regulate gene expression. The presence of promoter-proximal R-loops directly affects the binding of transcription factors, such as the Prc2 repressive complex and the Tip60–p400 histone acetyltransferase complex in mouse ES cells, which affects ES cells differentiation[110]. R-loops positioned near active CGI promoters also act as barriers against DNA methylation spreading, keeping CGIs in an unmethylated state likely by reducing the affinity of DNMTs binding to DNA[70,91] or by recruiting TET demethylation machinery[113]. Additionally, R-loops are linked to heterochromatin establishment, as observed over G-rich terminator elements, where R-loops promote H3K9me2, a repressive mark that reinforces RNA Pol II pausing during transcription termination[104].

1.4.3. R-loops as drivers of genomic instability

Although playing critical roles in vital cellular processes, R-loops need to be tightly controlled, or otherwise they pose severe threats to genome stability. The first experimental evidence of R-loop-driven DNA damage was observed in yeast mutants of the THO/TREX complex, which is involved in pre-mRNA export. These mutants present exacerbated R-loop levels and hyper-recombination phenotype, which can be reverted by R-loops removal[74]. Also, in yeast, loss of Sen1 helicase causes R-loop accumulation and the consequent transcription-associated recombination[98]. In human cells, the suppression of transcription, splicing or mRNA-processing factors elicits R-loop-dependent DNA damage. Examples of these are the SRSF1 splicing factor, AQR RNA helicase or SETX helicase[66].

Unresolved R-loops are a source of several genomic instability events. The thermodynamically stable conformation of R-loops makes their resolution an energy-consuming process[63]. Moreover, the displaced ssDNA is more susceptible to lesions, since it is the suitable substrate for DNA-modifying enzymes, such as AID. AID-mediated mismatches can lead to DNA single-strand breaks (SSBs)[66], which in turn may evolve into DNA DSBs, the most deleterious form of DNA damage[114]. ssDNA is also more susceptible to spontaneous mutagenicity. An example of that is the 140-fold higher spontaneous deamination of cytosine observed in ssDNA when compared to dsDNA[115]. Furthermore, the presence of RNA in the double helix is shown to inhibit nucleosome assembly, creating "naked" DNA regions more vulnerable to lesions[66,116]. Lastly, R-loops can stall the progression of the transcription machinery, posing a physical barrier to the replication fork and causing transcription-replication collisions. This leads to replicative stress and eventual fork collapse[66,117,118]. Such collisions, which have been described in bacteria, yeast and mammals[118], are the most common cause of R-loop-driven DNA damage and are extremely dangerous as they may result in DNA DSBs, recombination-mediated repair, chromosome rearrangements and ultimately cell death[119]. Nonetheless, replication-transcription encounters are functionally important in specific events, as immunoglobulin class switch recombination in vertebrate B cells, in which R-loop-mediated collisions allow the necessary DNA rearrangements[65,119].

Interestingly, R-loop formation as a consequence of DNA damage events has also been described. For instance, transcription-replication encounters and the consequent hindering of DNA polymerase progression may favor R-loop formation[119]. Moreover, induced DNA lesions were shown to cause RNA Pol II pausing and spliceosome displacement (required for subsequent DNA repair), resulting in the release of the pre-mRNA that becomes free to hybridize with the template ssDNA. Notably, these R-loops engage in DDR feedback, recruiting DNA repair factors to the damaged locus[120].

Intriguingly, there are some cases of regulatory R-loops that rarely lead to DSBs, such as those involved in bacterial and mitochondrial replication, gene expression, or DNA repair. Although this topic remains largely unexplored, the abundancy, the genomic context, and the persistence of such R-loops have been raised as putative explanations for their unusual outcome[66].

2. Aims

5hmC instructs mechanistic and thermodynamic changes in the DNA double helix that lead to a more open and relaxed chromatin state, increasing base-pair flexibility and reducing dsDNA melting temperature. Besides the direct impact on transcription factor binding and gene expression, this may also favor the invasion of the double helix by foreign molecules, such as RNA, leading to R-loop formation. Indeed, although there are studies revealing the influence of R-loops in chromatin remodeling through different epigenetic mechanisms, nothing is known about the crosstalk between 5hmC and R-loop structures. Therefore, we asked the following questions:

#1: Does transcription through 5hmC-rich templates favor R-loop formation?

Rationale: Since DNA features that destabilize the double helix, such as supercoiling or G-quadruplexes, are known to facilitate nascent RNA annealing with the template DNA strand, we reasoned that 5hmC may favor R-loop formation due to the above mentioned mechanistic and thermodynamic effects that this epigenetic mark exerts over dsDNA.

#2: How does TET enzymatic activity affect R-loop prevalence?

Rationale: Owing to TETs' role in converting 5mC into 5hmC, we hypothesized that editing 5hmC density either by changes in TETs expression levels or by targeting their enzymatic activity to specific loci will impact R-loop formation.

#3: What is the physiological relevance of a putative crosstalk between 5hmC and R-loops?

Rationale: In cellular settings that require high gene plasticity, as ES cells' fate commitment and lineage specification, profound changes in gene expression rely on epigenetic reprogramming, to which TET enzymes are crucial. Hence, we seek to characterize the 5hmC/ R-loop axis at the genome-wide level and to investigate the relevance of such crosstalk in ES cells-related processes. Also, we reason that 5hmC-mediated chromatin changes may render the dsDNA more vulnerable, and so we aim to explore 5hmC as a putative predictor of DNA damage events.

Goal

The main goal of this thesis is to provide insights into each of the aforementioned questions and to clarify the interplay between 5hmC and R-loop structures.

3. Materials and Methods

3.1. Cell lines and culture conditions

E14TG2a (E14) mouse ES cells were provided by Domingos Henrique (Instituto de Medicina Molecular João Lobo Antunes), and were a gift from Austin Smith (Univ. of Exeter, UK)[121]. 129S4/SvJae (J1) mouse ES 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 amino acids 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). The 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 (all purchased from ATCC) 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% CO_2.

3.2. *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 mouse ES cells stably expressing a doxycycline-inducible short hairpin RNA targeting *Tet3* (Supplementary Table 1). Cells were treated for 48h with 2 μg/mL doxycycline (D9891, Sigma Aldrich).

3.3. 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: $2^{(Ct\ reference\ -\ Ct\ sample)}$, where Ct reference and Ct sample are mean threshold cycles of RT-qPCR done in duplicate for *U6* snRNA or *Gapdh* mRNA and the gene of interest, respectively. Primer sequences are presented in Supplementary Table 2.

3.4. 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 37ºC, 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 60ºC, followed by blocking and re-probing. Details of antibodies used are included in Supplementary Table 3.

3.5. Proximity Ligation Assay (PLA)

E14 mouse ES 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 37ºC, 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 4′,6-diamidino-2-phenylindole (DAPI) in Vectashield. For the RNase H control, fixed cells were treated with 3U/μL RNase H (MB085, NZYTech) for 1h at 37ºC before 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 Table 3.

3.6. g-blocks PCR

Designed g-blocks were ordered from IDT (Supplementary Table 4), 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 Table 2), 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 Table 3.

3.7. *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. For RNA recovery, the resulting RNA was column-purified with NucleoSpin RNA isolation kit (740955.250, Macherey-Nagel) and quantified in a NanoDrop 2000 spectrophotometer (Thermo Scientific).

3.8. Dot Blot of R-loops formed *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 an 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 PBSt 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 Table 3.

3.9. Radioactive labelling of *in vitro* transcription templates

Membranes were incubated in denaturing solution (1,5M NaCl, 0,5M NaOH) for 30min and then in neutralization solution (0,5M Tris-HCl pH 7.5, 3M NaCl) for 15min, at room temperature. Membranes were subsequently pre-hybridized in Church buffer (0,25M sodium phosphate buffer pH 7.2, 1mM EDTA, 1% BSA, 7% SDS) at 50°C for 2h and then hybridised overnight at 50°C with an oligonucleotide probe (Supplementary Table), 5'-end labelled with T4 polynucleotide kinase (New England Biolabs) and [γ-32P]ATP. Post-hybridisation washes were performed twice in 2x SSC (0,3M NaCl, 0,03M trisodium citrate pH 7.2), 0,2% SDS for 20min and once in 1x SSC, 0,2% SDS for 30min, at 50°C. Storage phosphor screens were exposed to the membranes and radioactive signals were detected using a Typhoon FLA 9000 imager (GE Healthcare). Signal quantification was performed using ImageJ.

3.10. 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. The DNA solution was diluted 1:10 in Sigma ultrapure water (with final 10mM $MgCl_2$) 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 the 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.

3.11. CRISPR-assisted 5mC/5hmC genome editing

Lentiviruses 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. Lentiviruses were collected through ultracentrifugation (25000 rpm, 3h, 4°C) using an SW-41Ti rotor in a Beckman XL-90 ultracentrifuge. Viruses 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.

3.12. 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 XhoI), 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 Na_2HPO_4 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: $2^{(Ct\ Input - Ct\ IP)}$, 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 Table 2**.**

3.13. 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 each 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 Na_2HPO_4 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: $2^{(Ct\ Input - Ct\ IP)}$, 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 Table 2.

3.14. Cell cycle analysis

pEGFP-N1 (GFP coding plasmid used as control) was purchased from Addgene, and pEGFP-RNaseH1 (GFP-tagged RNase H1 coding plasmid) was kindly provided by Robert J. Crouch (NIH, USA). Seeded mouse ES 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.

3.15. 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 Table 5)**.** 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 run 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.

3.16. Multi-omics data

High-throughput sequencing (HTS) data for mouse ES cells and HEK293 cells were gathered from GEO archive: transcriptome of mouse ES cells (GSE67583); R-loops in mouse ES cells (GSE67581); 5hmC in mouse ES cells (GSE31343); ɣH2AX in mouse ES cells (GSE69140); active transcription in HEK293 (GRO-seq, GSE51633); R-loops in HEK293 (DRIP-seq, GSE68948); 5hmC modification in HEK293 (hMeDIP-seq, GSE44036); ɣH2AX (ChIP-seq, GSE75170). Transcriptome profiles of mouse ES cells overexpressing RNase H were obtained from GSE67583. The quality of HTS data was assessed with FastQC (www.bioinformatics.babraham.ac.uk/projects/fastqc).

3.17. 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[122], and filtering for uniquely aligned reads. Enriched regions were identified relative to the input samples using MACS[123], with a false-discovery rate of 0,05. Finally, enriched regions were assigned to annotated genes, including a 4 kbp region upstream of the TSS and downstream of the TTS. 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[124]. For individual and metaprofiles, uniquely mapped reads were extended in the 3' direction to reach 150 nucleotides with the Pyicos[125]. Individual profiles were produced using a 20bp window. For the metaprofiles centred 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 was 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 R environmental language http://www.R-project.org[126]. SAMtools[127] 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 original datasets). The p-value was determined as the frequency of overlapping regions between the random datasets as extreme as the observed.

3.18. Transcriptome analysis

Expression levels (Transcripts per Million, TPMs) from RNA-seq and GRO-seq datasets were obtained using Kallisto[128], 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 25th percentile. Differential expression in mouse ES cells overexpressing RNase H was assessed using edgeR (v3.20.9) and limma (v3.34.9) R packages[129,130]. 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)[131] and filtered using False discovery rate corrected p-values < 0.05.

For the analysis of transcription readthrough, transcriptome profiles from human ES cells (WT and *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[132]. Transcription readthrough levels were evaluated by counting the number of reads mapping downstream the TSS using ARTDeco[133] and human genome annotation from the GENCODE project (GENCODE release 37). Genes with enrichment in transcriptional readthrough in *TET1* KO samples relative to the control were identified. Metagene profiles were built using the *computeMatrix* tool from the deepTools v3.5.1[134] 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 1 kbp 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).

4. Results

The results presented below were compiled and published in eLife peer-reviewed journal, under the title:

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

The article, in the publication form, is included in the Annexes.

Authors: João C. Sabino[1], Madalena R. de Almeida[1], Patrícia L. Abreu[1], Ana M. Ferreira[1], Paulo Caldas[2,3], Marco M. Domingues[1], Nuno C. Santos[1], Claus M. Azzalin[1], Ana R. Grosso[2,3], Sérgio F. de Almeida[1*]

Affiliations: [1]Instituto de Medicina Molecular João Lobo Antunes, Faculdade de Medicina da Universidade de Lisboa, Lisboa, Portugal.
[2]Associate laboratory i4HB – Institute for Health and Bioeconomy, NOVA School of Science and Technology, Universidade Nova de Lisboa, Caparica, Portugal.
[3]UCIBIO-REQUIMTE, Applied Molecular Biosciences Unit, Department of Life Sciences, NOVA School of Science and Technology, Universidade Nova de Lisboa, Caparica, Portugal.
*Corresponding author

DOI: 10.7554/eLife.69476

Author contributions:

João Sabino: Data curation; Conceptualization and design; Data analysis and interpretation; Drafting or revising the article. Sérgio de Almeida: Conceptualization and design; Data analysis and interpretation; Supervision; Funding acquisition; Drafting or revising the article. Madalena de Almeida: Data curation; Data analysis and interpretation. Patrícia Abreu: Data curation; Data analysis and interpretation. Ana Ferreira: Data analysis and interpretation. Paulo Caldas: Data analysis and interpretation. Marco Domingues: Data curation; Data analysis and interpretation; Drafting or revising the article. Nuno Santos: Data analysis and interpretation; Drafting or revising the article. Claus Azzalin: Data analysis and interpretation; Drafting or revising the article. Ana Grosso: Software; Data analysis and interpretation; Drafting or revising the article.

4.1. Transcription through 5hmC-rich DNA favors R-loop formation

To assess the impact of cytosine (hydroxyl)methylation on R-loop formation, we performed *in vitro* 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[101,104]. Two of these sequences (*β-actin* P1 and *β-actin* P2) are from the transcription termination region of the *β-actin* gene, while the third sequence is from the *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 6A). Successful incorporation of dCTP variants was confirmed by immunoblotting using specific antibodies against 5mC and 5hmC variants (Figure 6B). *In vitro* transcription was initiated from the embedded T7 promoter sequence, and the formation of R-loops during transcription of each template was inspected. To this end, *in vitro* reaction products were run in an agarose gel, immobilized on a nylon membrane through capillary transfer, and immunoblotted with the S9.6 antibody (S9.6 Ab), which binds DNA:RNA hybrids (Figure 6C-upper panel). 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 6C-upper panel). R-loop signal was normalized to the amount of DNA template in each condition, assessed with a radioactively labelled oligonucleotide probe (Figure 6C-lower), and plotted. In agreement with our hypothesis that 5hmC favors R-loops, increased amounts of DNA:RNA hybrids were detected in samples derived from *in vitro* transcription of 5hmC-rich *β-actin* P1, *β-actin* P2 and *APOE* DNA templates (Figure 6D).

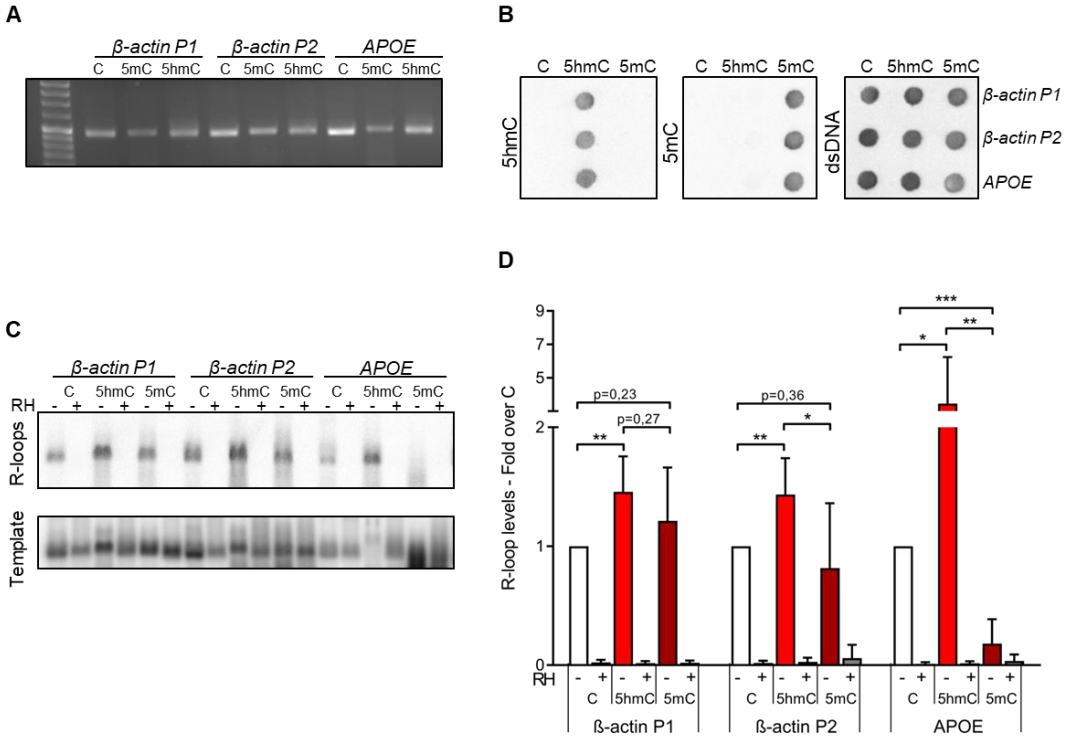


Figure 6: 5hmC favors co-transcriptional R-loop formation. (A) Native or modified dCTPs were incorporated upon PCR amplification into DNA fragments with sequences from the transcription termination region of the *β-actin* gene (*β-actin* P1 and *β-actin* P2) or the *APOE* gene. (B) Incorporation of dCTP variants confirmed by immunoblotting using specific antibodies against 5mC, 5hmC and 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.

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, 5hmC, or 5mC. DNA:RNA hybrids formation by annealing of ssDNA with the complementary single-stranded RNA (ssRNA) was confirmed through band shift in a polyacrylamide gel (Figure 7A). EMSA of DNA:RNA hybrids incubated with increasing amounts of the S9.6 Ab revealed that S9.6 Ab was able to delay the run of the three substrates with similar kinetics (Figure 7B). Quantification of free hybrids in the native polyacrylamide gel endorses S9.6 capacity to equally recognize DNA:RNA hybrids formed with any of the three C variants (Figure 7C).

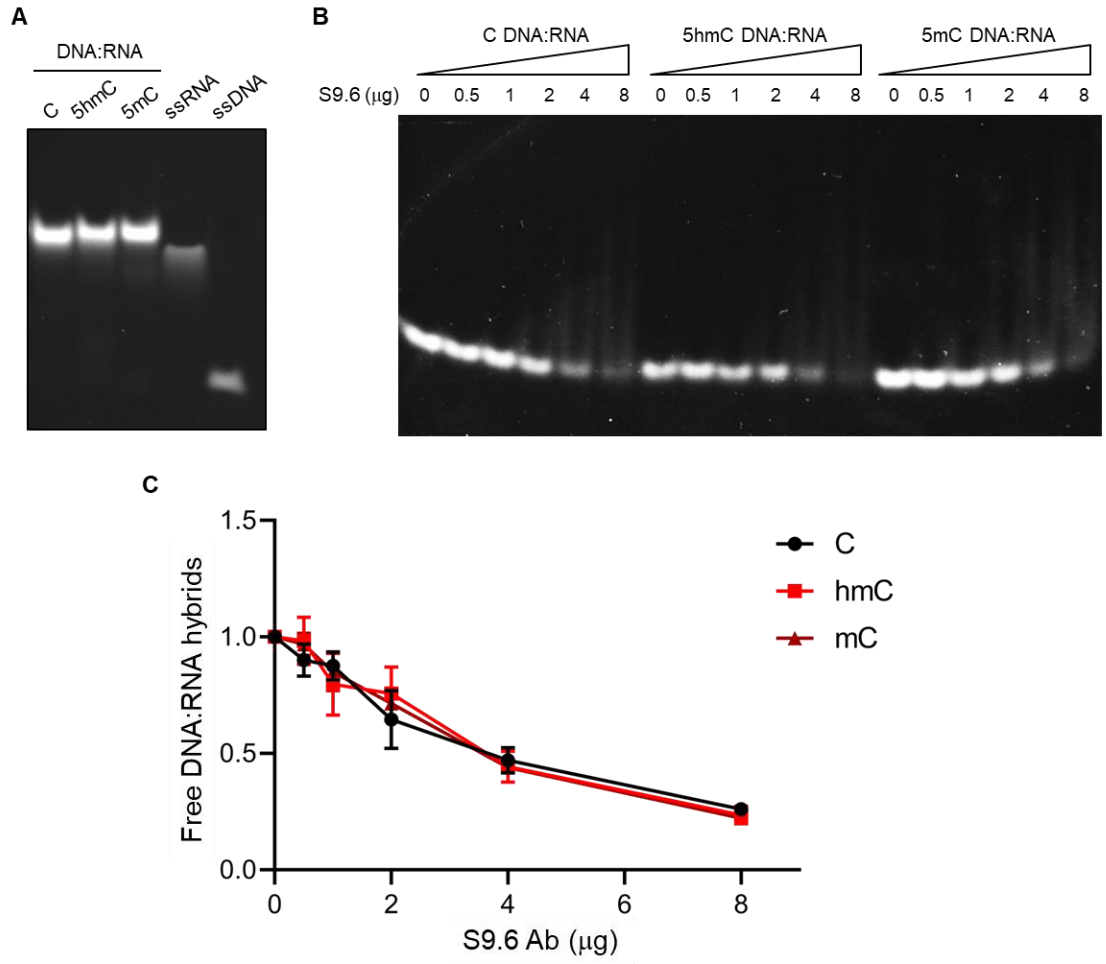


Figure 7: Cytosine modifications do not affect the detection of DNA:RNA hybrids by the S9.6 Ab. (A) DNA:RNA hybrids formed through annealing of C, 5hmC or 5mC-containing ssDNA with the complementary ssRNA. Data are representative of three independent experiments. (B) EMSA of DNA:RNA hybrids incubated 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.

We next sought to directly visualize R-loop structures obtained in the *in vitro* transcription reactions through atomic force microscopy (AFM). *β-actin* P2 transcription products were treated with Proteinase K to remove transcriptional machinery and improve nucleic acids observation. Figure 8A shows field acquisitions of C, 5hmC and 5mC transcription products, either treated or not with RNase H. In a first analysis, RNase H treated samples depict longer filaments, likely as a result of topological constraints removal due to the elimination of hybrids. R-loops were identified as previously described[135,136], and each individual DNA molecule establishing an R-loop structure in the AFM images was assigned manually. The frequency of these structures, which are extensively lost upon RNase H treatment (Figure 8B), 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

8C). 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.

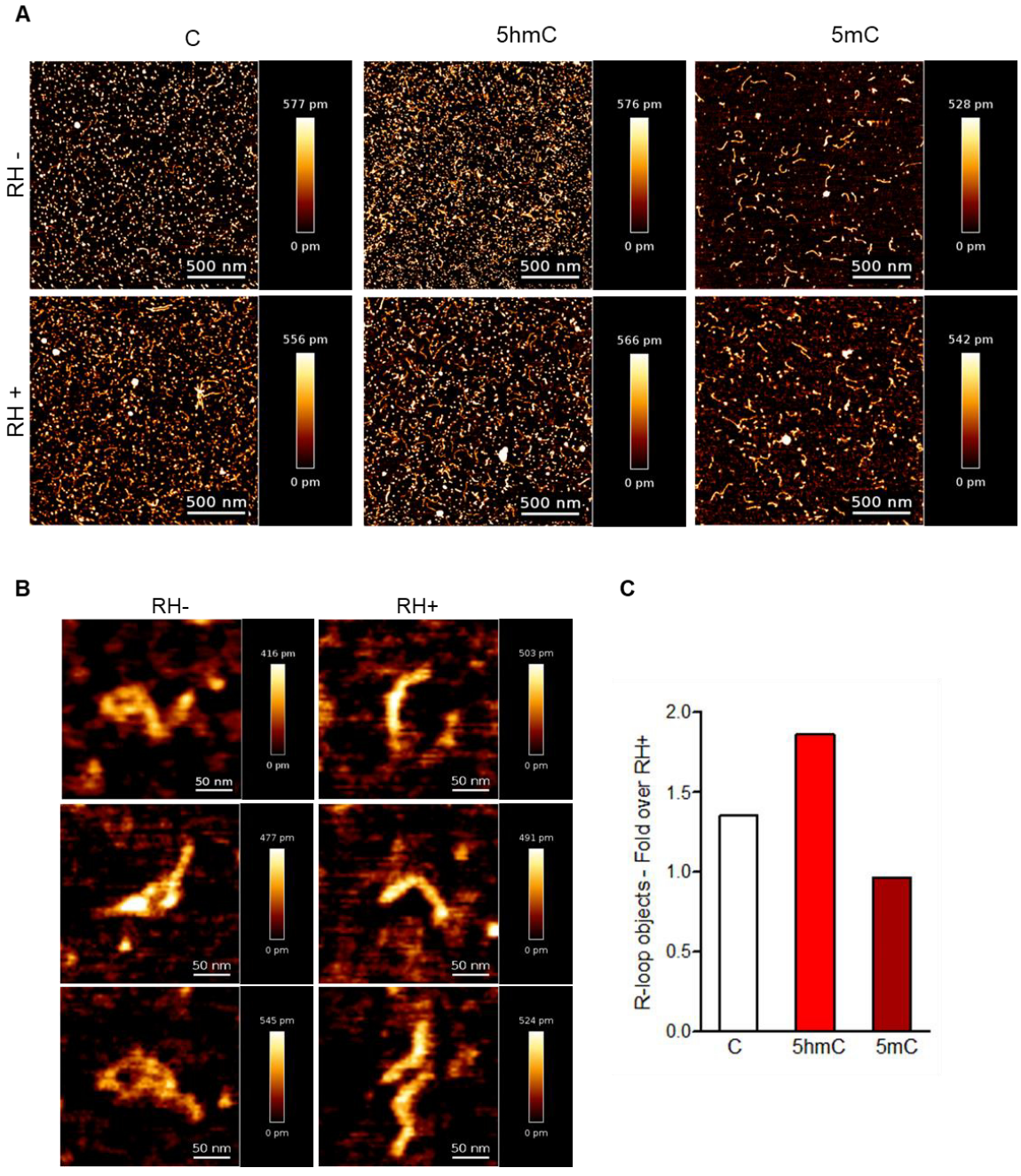


Figure 8: *In vitro* transcription reaction products of *β-actin* P2 templates were visualized using AFM. (A) Field acquisitions of C, 5hmC and 5mC transcription products, in the absence (RH-) or presence (RH+) of RNase H. Scale bars: 500nm. (B) R-loop structures obtained from 5hmC-containing *β-actin* P2 transcription, which are extensively lost upon RNase H treatment. (C) R-loops present in the transcription reaction products of C, 5mC or 5hmC-containing *β-actin* P2 templates were counted in a minimum of 80 filaments observed in three individual AFM experiments.

Moreover, we wanted to investigate if DNA modifications impact transcription levels of our *in vitro* model. For that, we column-purified and quantified the RNA synthesized upon transcription of DNA templates containing unmodified C, 5hmC or 5mC (Figure 9A). 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 *in vitro* (Figure 9B). On one side, higher R-loop detection on a diminished transcription setting, as in the case of 5hmC templates, further strengthens our hypothesis that 5hmC favors R-loop formation. However, we cannot conclude about the impact of 5mC modification 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 *in vivo.*

A

DNA template	RNA concentration
β-actin P1 C	414,40 ng/uL
β-actin P1 5hmC	271,00 ng/uL
β-actin P1 5mC	71,17 ng/uL
β-actin P2 C	128,77 ng/uL
β-actin P2 5hmC	40,30 ng/uL
β-actin P2 5mC	28,60 ng/uL
APOE C	188,63 ng/uL
APOE 5hmC	108,47 ng/uL
APOE 5mC	22,23 ng/uL

B

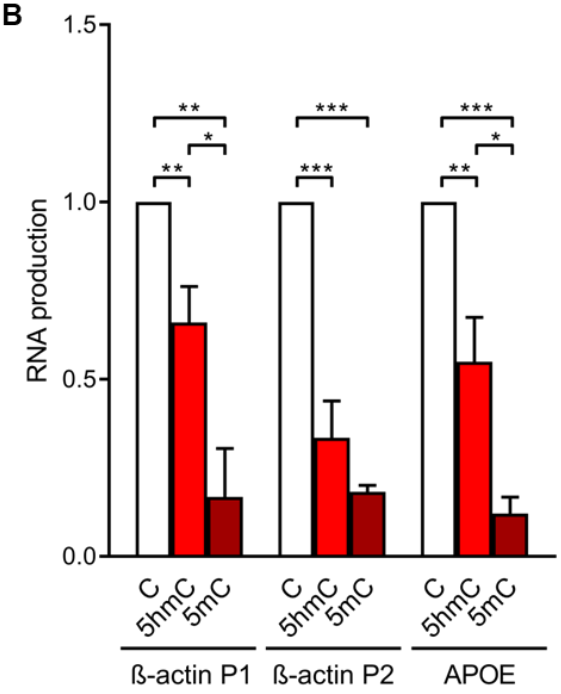


Figure 9: *In vitro* transcription levels of different DNA templates. (A) RNA concentration obtained from *in vitro* transcription of DNA templates containing C, 5hmC, or 5mC. Data shown are the mean from three independent experiments. (B) Mean and SD of the RNA levels obtained in each reaction normalized to reactions using DNA templates with native cytosines. Data are from three independent experiments.

4.2. Editing 5hmC density impacts endogenous R-loops

4.2.1. Changes in *Tet* expression levels influence R-loop formation

To test whether the 5hmC DNA modification induces R-loop formation *in vivo*, we quantified R-loop levels in RNAi-mediated *Tet*-depleted mouse ES cells. Since Tet enzymes are responsible for oxidizing 5mC into 5hmC, we expected to reduce global 5hmC levels upon Tets knockdown. Despite the significant reduction in *Tet1*, *Tet2* and *Tet3* expression (Figure 10A), the levels of 5hmC were not significantly affected by *Tet1* or *Tet2* depletion, as observed by immunoblotting of total cellular nucleic acids (Figure 10B,C). 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 10B,C). This finding suggests that there is a partial redundancy in the activity of the three Tet enzymes in mouse ES cells. The loss of Tet1 or Tet2 - but not of Tet3 - is compensated by the remaining Tets. Unsurprisingly, no significant changes were observed in 5mC levels (Figure 10B,C), as 5hmC is typically 10–100 times less represented in the genome than 5mC[50].

In agreement with the hypothesis that 5hmC promotes R-loop formation, dot-blot hybridization of total cellular nucleic acids using the S9.6 Ab also revealed significantly reduced endogenous R-loop levels in mouse ES cells after depletion of Tet3 and after co-depletion of the three Tet enzymes (Figure 10B,D). Indeed, 5hmC and R-loop dynamics upon Tets depletion are remarkably synchronized.

Besides the results obtained in mouse ES cells, we also measured R-loop levels upon RNAi depletion of Tet enzymes in NIH-3T3 mouse fibroblasts. Although *Tet1*, *Tet2* and *Tet3* expression levels were significantly diminished (Figure 11A), a significant reduction of 5hmC was only obtained upon depletion of Tet3 and of the three Tets in mouse fibroblasts (Figure 11B,C), as observed in mouse ES cells, pointing towards the Tets redundancy mechanism previously mentioned. Also, no meaningful changes in 5mC were detected (Figure 11B,C). Regarding R-loops, *Tets* triple knockdown significantly reduced the global levels of R-loops in mouse fibroblasts, whereas Tet3 depletion in these cells had a minor impact (Figure 11B,D).

Next, we wanted to further confirm this effect at the single gene level. For that, we measured R-loops formed at selected active genes by DNA:RNA immunoprecipitation (DRIP) in mouse ES cells (Figure 10E) and in mouse fibroblasts (Figure 11E), following

Tet1/2/3 triple knockdown. The DRIP assays confirmed that R-loops are less abundant in the analysed genes upon depletion of Tet enzymes.

Given the association of 5hmC with genes' active transcription state, we then asked whether transcription rate changes in Tet1/2/3-depleted cells could contribute to the diminished R-loop formation. We observed that simultaneous depletion of the three enzymes did not affect the expression levels of the analysed genes in both mouse ES cells (Figure 10F) and mouse fibroblasts (Figure 11F). Therefore, these data suggest that the activity of Tet enzymes promotes the formation of R-loops in absence of changes in transcription levels.

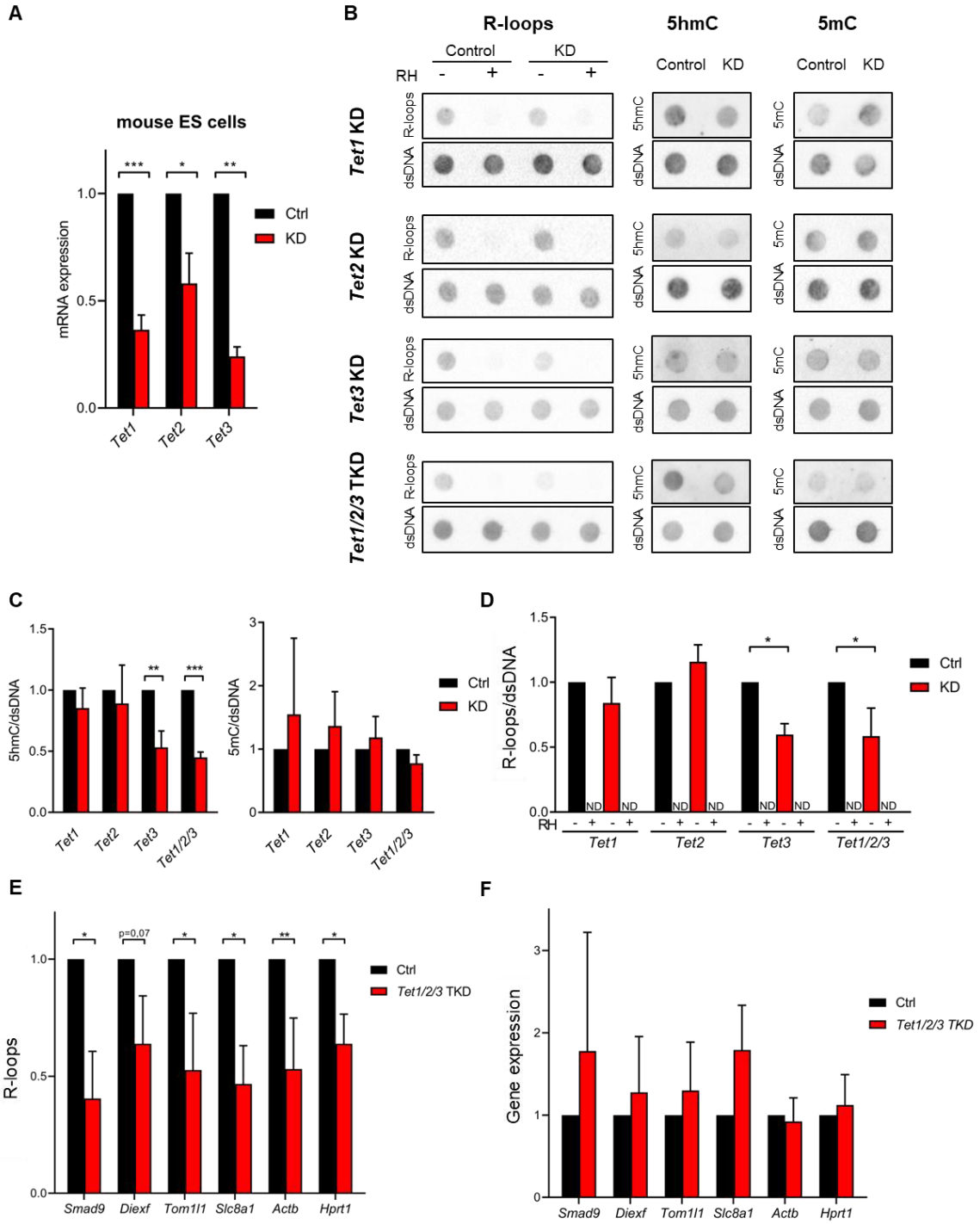


Figure 10: *Tet* depletion impacts R-loop formation in mouse ES cells. (A) *Tet1, Tet2* and *Tet3* mRNA expression levels in mouse ES cells 48h after RNAi depletion of *Tet1*, *Tet2* or *Tet3*. *p<0.05, **p<0.01, ***p<0.001, two-tailed Student's t test. (B) Dot blot of R-loops, 5hmC and 5mC in *Tet1*, *Tet2* and *Tet3* single KD and in *Tet1/2/3* triple KD mouse ES cell extracts. dsDNA was detected after stripping and re-probing of the same membranes. Data are representative of a minimum of three independent experiments. Quantification of 5hmC and 5mC (C) and R-loops (D) dot blots shown in (B). Data were normalized against dsDNA levels. *p<0.05, **p<0.01, ***p<0.001, two-tailed Student's t test. ND=not detected. (E) R-loop levels assessed by DRIP, in *Tet1/2/3* triple KD mouse ES cells. Data were normalized against RNase H-treated samples. *p<0.05, **p<0.01, two-tailed Student's t test. (F) Transcription levels of the genes presented in (E) assessed by RT-qPCR. All graphs show the mean and SD from a minimum of three independent experiments.

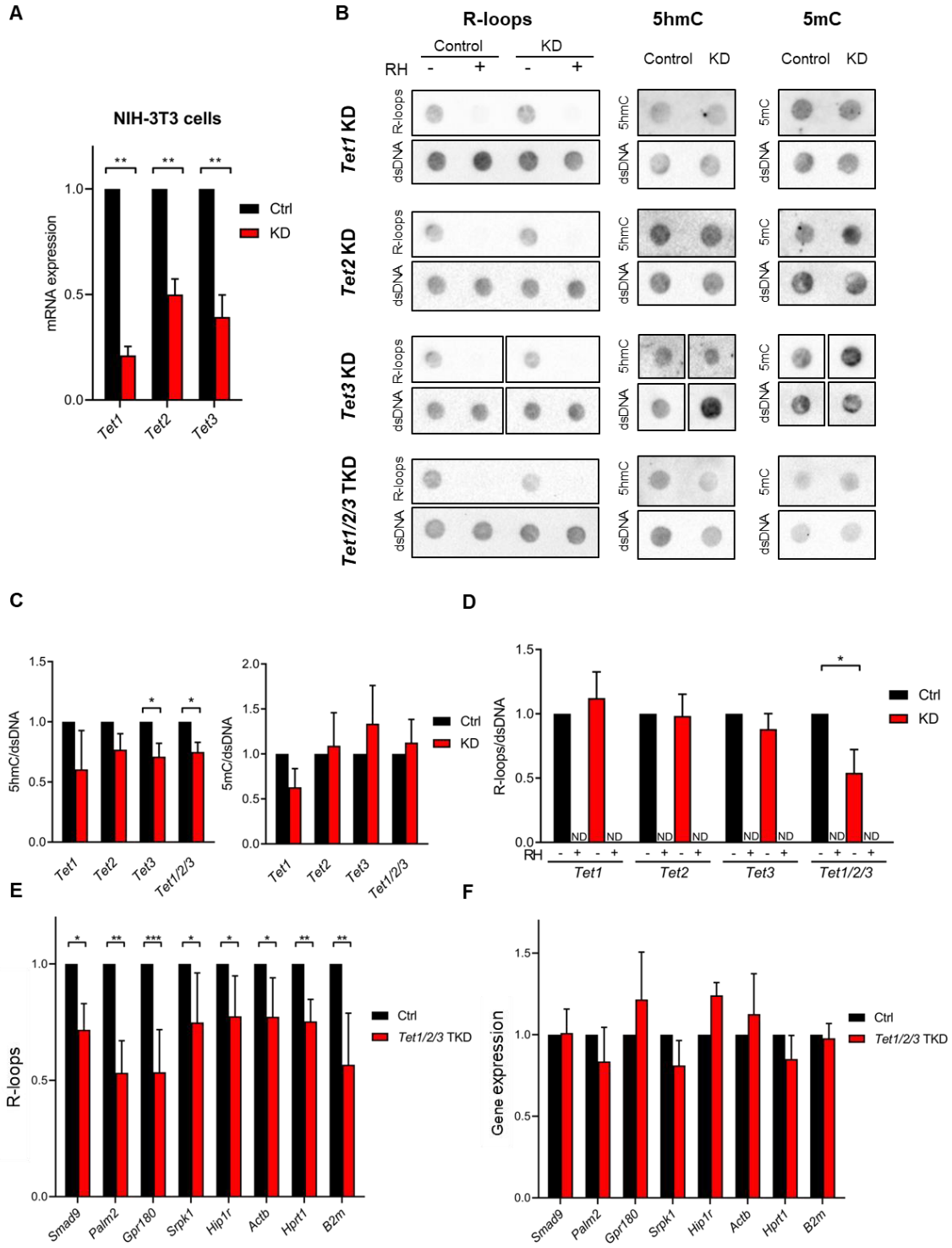


Figure 11: *Tet* depletion impacts R-loop formation in mouse fibroblasts. (A) *Tet1, Tet2* and *Tet3* mRNA expression levels in mouse fibroblasts 48h after RNAi depletion of *Tet1*, *Tet2* or *Tet3*. **p<0.01, two-tailed Student's t test. (B) Dot blot of R-loops, 5hmC and 5mC in *Tet1*, *Tet2* and *Tet3* 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 representative of a minimum of three independent experiments. Quantification of 5hmC and 5mC (C) and R-loops (D) dot blots shown in (B). Data were normalized against dsDNA levels. *p<0.05, two-tailed Student's t test. ND=not detected. (E) R-loop levels assessed by DRIP, in *Tet1/2/3* triple KD mouse fibroblasts. Data were normalized against RNase H-treated samples. *p<0.05, **p<0.01, ***p<0.001, two-tailed Student's t test. (F) Transcription levels of the genes presented in (E) assessed by RT-qPCR. All graphs show the mean and SD from a minimum of three independent experiments.

4.2.2. Targeted Tet enzymatic activity promotes R-loop formation

To further validate the impact of 5hmC in R-loop formation, we employed a modified CRISPR-based system to target TET enzymatic activity to specific loci[137]. We used a pool of three specific guide RNAs (gRNAs) to direct a catalytically inactive Cas9 (dCas9) nuclease fused to the catalytic domain of TET1 (dCas9-TET) to the last exon of the *APOE* gene in human osteosarcoma (U-2 OS) cells. According to reduced representation bisulfite sequencing data from Ensembl Genome Browser (http://.ensembl.org/), the target region contains several methylated CpGs. As a control, dCas9 was fused to an inactive mutant version of the TET1 catalytic domain (dCas9-dTET).

Constructs and gRNAs were introduced in the cells through viral delivery, and local enrichment of 5hmC following dCas9-TET targeting at the *APOE* locus was confirmed by DNA immunoprecipitation using antibodies specific for 5mC or 5hmC modified nucleotides – 5-(hydroxy)Methylated DNA Immunoprecipitation ((h)MeDIP). Indeed, the highest and most significant 5hmC/5mC ratio was detected at the gene segment adjacent to the gRNAs-target region (Figure 12A). Also, R-loop levels detected by DRIP peaked significantly at the gRNAs-target and in the downstream region upon tethering of dCas9-TET when compared to dCas9-dTET control (Figure 12B). Importantly, R-loop differences were not caused by changes in *APOE* gene expression levels, as they remain unchanged (Figure 12C). The increased levels of R-loops detected far from the dCas9-TET target site are consistent with the view that R-loops can 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 genes[138,139]. These results suggest that localized 5hmC enrichment can be sufficient to promote R-loop formation.

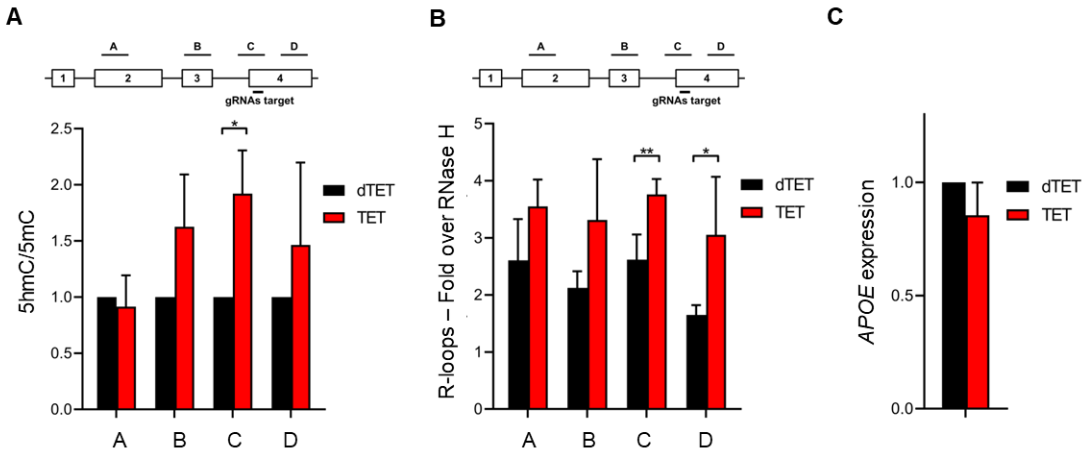


Figure 12: CRISPR-based TET tethering induces R-loop formation. 5hmC/5mC (A) and R-loop (B) levels determined by (h)MeDIP or DRIP at 4 regions of the *APOE* gene upon tethering of dCas9-TET1 or dCas9-dTET1 to the last exon of *APOE* in U-2 OS cells. R-loop data were normalized against RNase H-treated samples. *p<0.05, **p<0.01, two-tailed Student's t test. (C) *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.

4.3. 5hmC and R-loops overlap at transcriptionally active genes

4.3.1. 5hmC and R-loops overlap genome-wide

To further inspect the link between 5hmC and R-loops at the genome-wide level, we performed computational analyses of hMeDIP-seq and DRIP-seq datasets from mouse ES and HEK293 cells[110,140–142]. To assess individual genome-wide distribution profiles, 5hmC-rich regions were selected and aligned to their peak summit, and R-loops density was probed over fixed windows of ±10 kbp around the 5hmC peaks (Figure 13A and Figure 14A). The resulting metagene plots and heatmaps revealed a marked overlap between 5hmC-rich loci and R-loops, suggesting a strong co-localization of both structures.

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 kbp[110]), we could obtain a statistically significant Pearson correlation coefficient between both ($p<0.05$) (Figure 13B and Figure 14B). Furthermore, approximately half of all R-loops detected genome-wide in mouse ES cells, or one-third in HEK293 cells, occurred at 5hmC-containing loci (Figure 13C and Figure 14C). Notably, we

observed an overlap between 5hmC and R-loops in 51% of all actively expressed genes in mouse ES cells (Figure 13D).

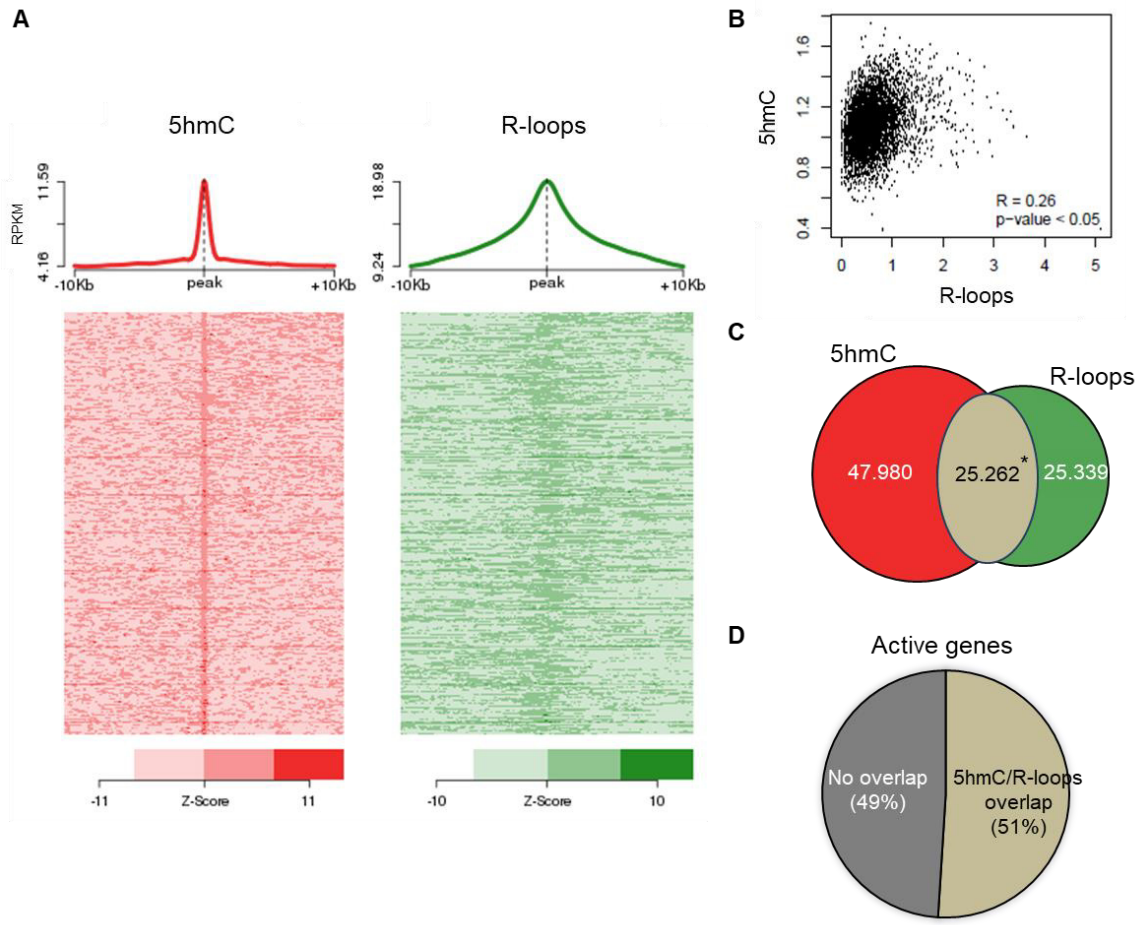


Figure 13: 5hmC and R-loops overlap in active genes of mouse ES cells. (A) Metagene and heatmap profiles of 5hmC and R-loops probed over fixed windows of ±10 kbp around the 5hmC peaks in expressed genes. (B) Pearson correlation coefficient between 5hmC and R-loops distribution within active genes ($p<0.05$). (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.

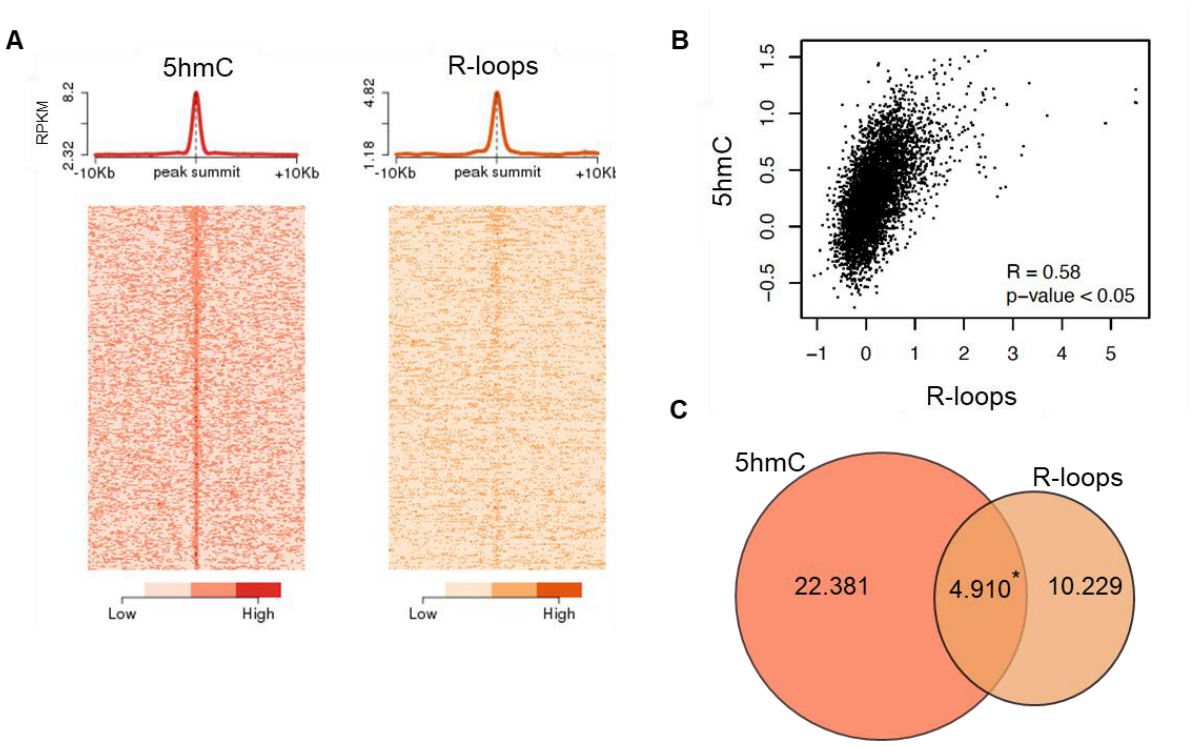
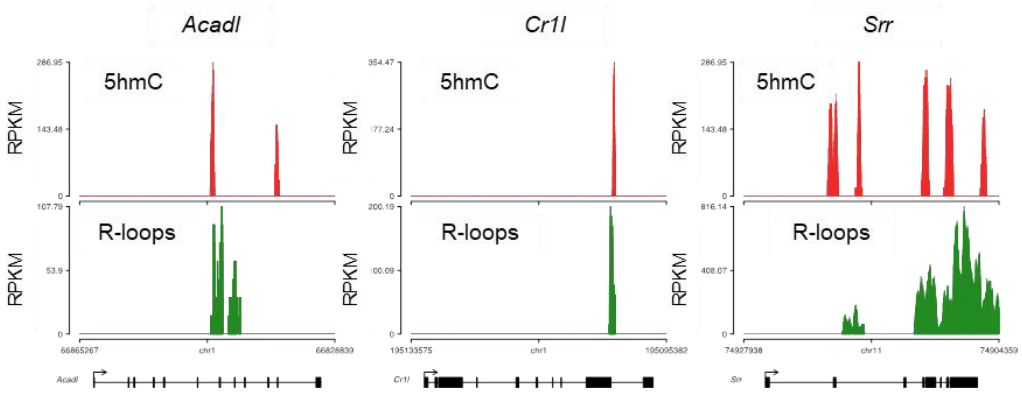


Figure 14: 5hmC and R-loops overlap in active genes of HEK293 cells. (A) Metagene and heatmap profiles of 5hmC and R-loops probed over fixed windows of ±10 kbp around the 5hmC peaks in expressed genes. (B) Pearson correlation coefficient between 5hmC and R-loops distribution within active genes ($p<0.05$). (C) Number of loci displaying 5hmC, R-loops, and overlapping 5hmC and R-loops. *Permutation analysis, $p<0.05$.

The individual profiles of 5hmC and R-loops in three mouse (Figure 15A) and human (Figure 16) expressed genes are well illustrative of such overlap. This feature is also evident in the individual distribution profiles of 5hmC and R-loops along two long regions of chromosome 17 (Figure 15B).

Altogether, these evidences strongly support 5hmC and R-loop co-localization at the genome-wide level, and thus stress out the view of R-loops' propensity to form in 5hmC-rich regions.

A


Acadl
Cr1l
Srr
5hmC
R-loops
RPKM

B

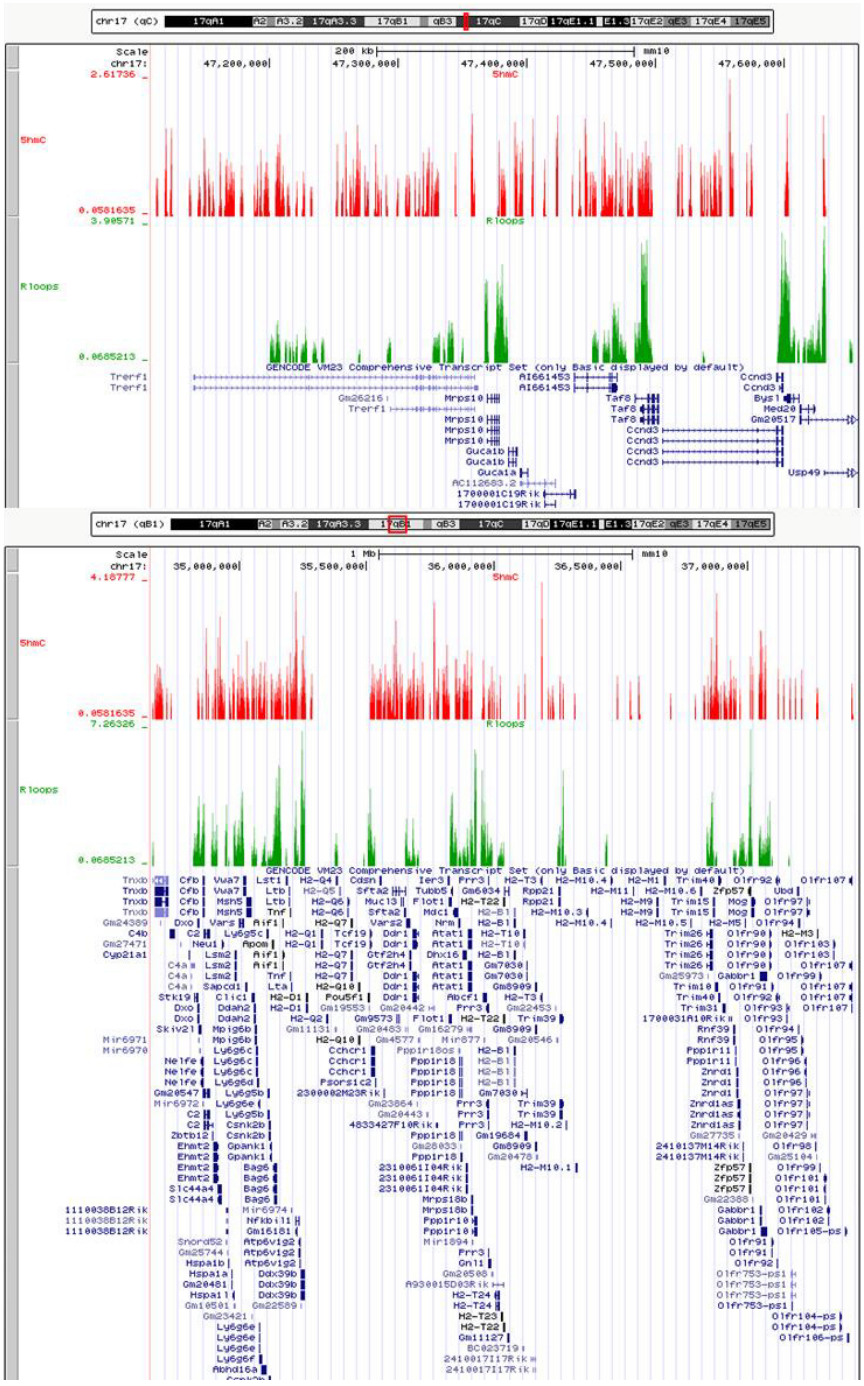

5hmC
R loops
GENCODE VM23 Comprehensive Transcript Set (only Basic displayed by default)

Figure 15: Individual genes and chromosomal distribution of 5hmC and R-loops in mouse ES cells. (A) Individual profiles of 5hmC and R-loop distribution along the *Acadl*, *Cr1l* and *Srr* genes. Density signals are represented as reads per kilobase per million mapped reads (RPKMs). (B) Individual profiles of 5hmC and R-loops along two long chromosomal regions. Density signals are represented as RPKMs and were uploaded in UCSC Genome Browser (http://genome-euro.ucsc.edu/).

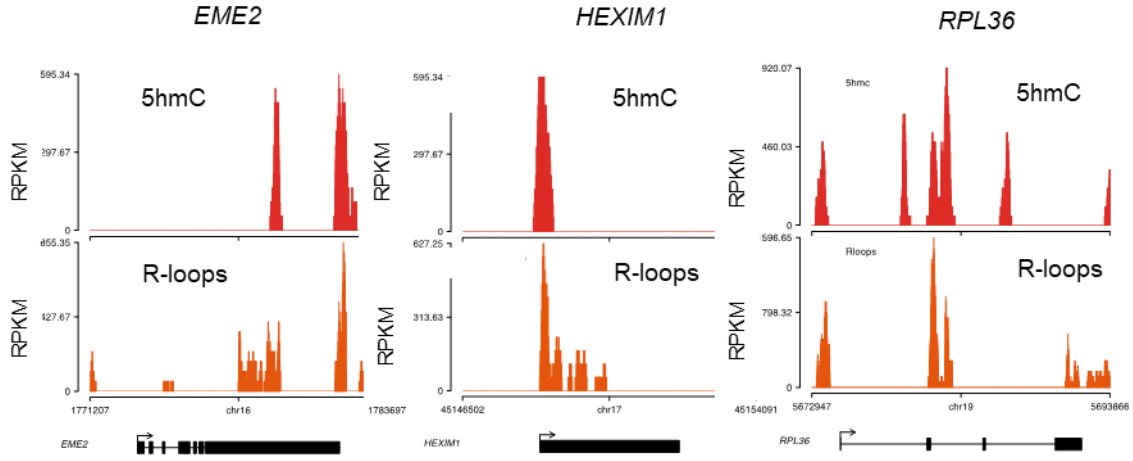


Figure 16: Individual genes distribution of 5hmC and R-loops in HEK293 cells. Individual profiles of 5hmC and R-loop distribution along the *EME2*, *HEXIM1* and *RPL36* genes. Density signals are represented as RPKMs.

4.3.2. 5hmC and R-loops are strongly correlated at the TTS

To better characterize 5hmC and R-loop overlapping in the active genome, metagene profiles of 5hmC and R-loops were generated (Figure 17A). Data revealed very similar patterns of intragenic distribution, with both 5hmC and R-loops increasing along the gene body towards the transcription termination site (TTS), where they reached maximum levels. At the TSS, however, the 5hmC DNA modification was mostly absent, whereas R-loops were abundant. Indeed, it is mainly in the gene body and termination region that the two structures overlap (Figure 17B). The detection of R-loop peaks at TSS regions is in agreement with previous studies[70,91] and implies that 5hmC is not necessary for co-transcriptional DNA:RNA hybridization and R-loop formation.

Interestingly, 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[143]. 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 17C). These data support a model whereby TET enzymes act upstream of R-loop formation during efficient transcription termination.

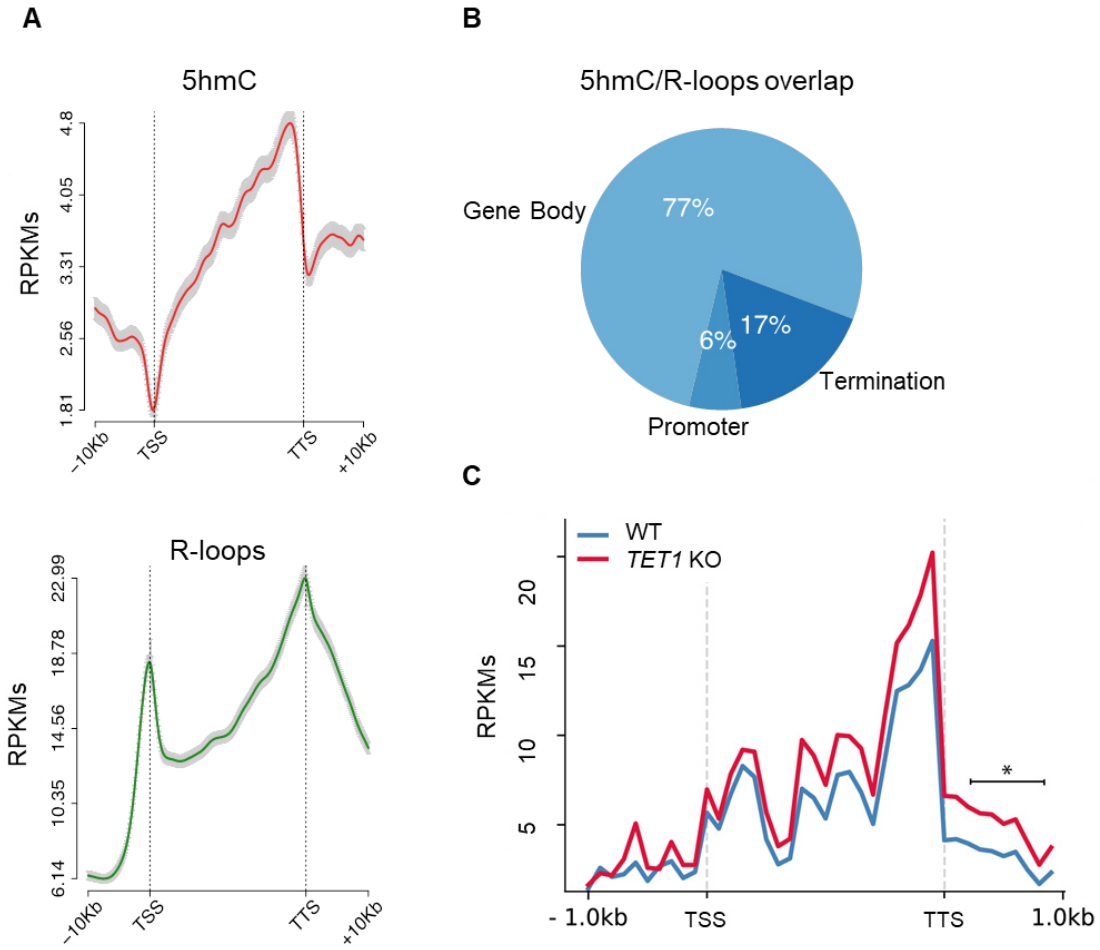


Figure 17: 5hmC and R-loops robustly overlap towards the TTS of active genes. (A) Metagene profiles of 5hmC and R-loop distribution in active genes. The gene body region was scaled to 60 equally-sized bins and ±10 kbp gene-flanking regions were averaged in 200 bp windows. Density signals are represented as RPKMs and error bars (gray) represent the standard error of the mean. (B) Percentage of 5hmC and R-loop overlapping events occurring in promoters, gene bodies and termination regions of expressed genes. (C) 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 into 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.

4.3.3. 5hmC and R-loops co-localize at the single-molecule level

Since 5hmC and R-loop structures are formed and actively resolved throughout cell cycle[34,61], we then sought to demonstrate 5hmC and R-loop simultaneous detection in individual mouse ES cells. We performed proximity ligation assays (PLA), a technique that allows detection and visualization of single events, such as molecule-to-molecule interactions in close proximity. We used S9.6 and anti-5hmC antibodies (Figure 18A). While control reactions without primary antibodies (-) and with each antibody alone did not produce a significant signal, staining of mouse ES cells with S9.6 and anti-5hmC antibodies gave rise to a robust PLA signal scattered throughout cells nuclei. Importantly, PLA signal was mostly lost after digestion of cells with RNase H, as reflected by the significant reduction in PLA foci per nucleus (Figure 18B), revealing the specificity of the PLA signal observed. These data confirm 5hmC and R-loop overlapping at the single-molecule level

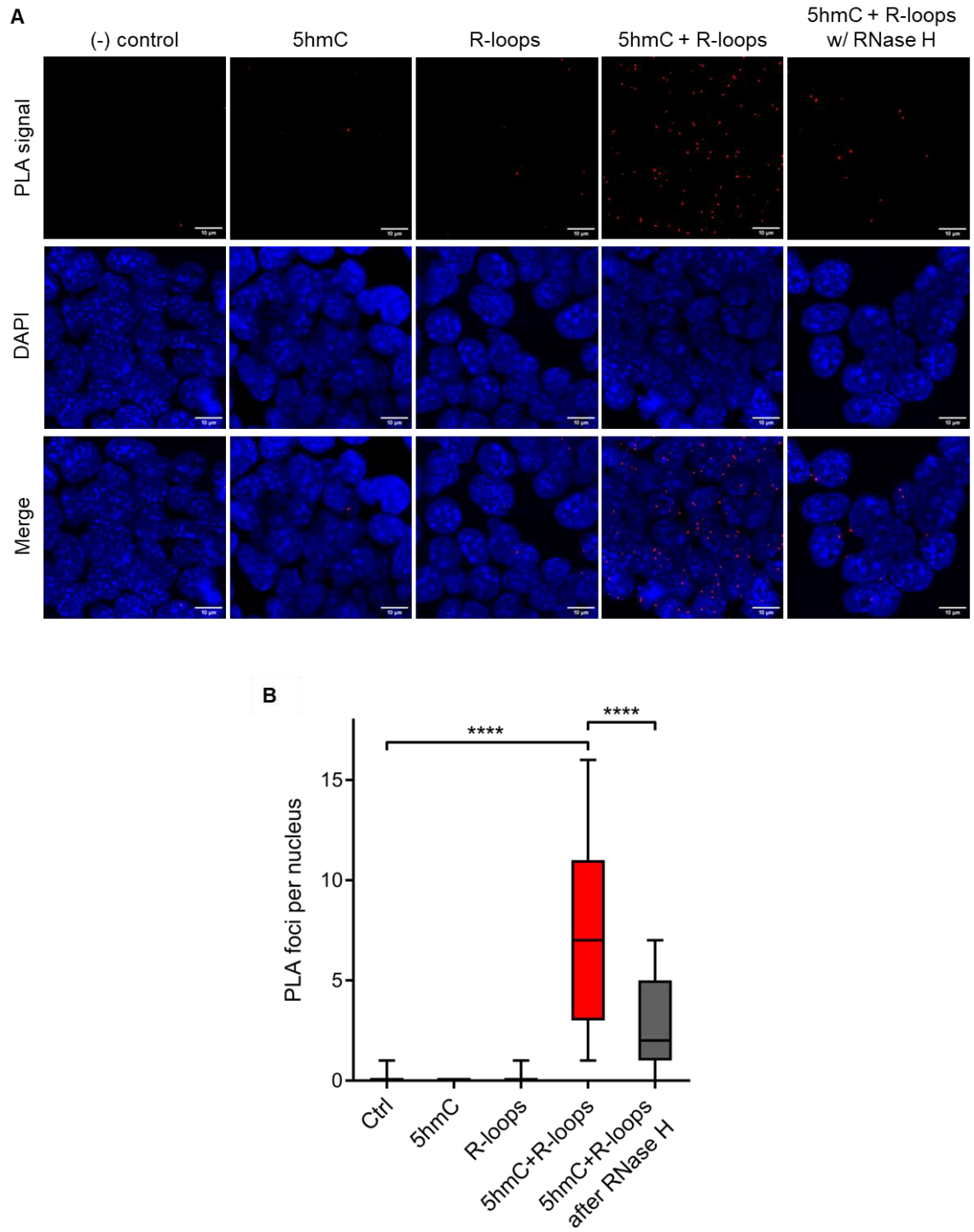


Figure 18: Simultaneous detection of 5hmC and R-loops at the same genomic loci in individual mouse ES cells. (A) 5hmC and R-loops PLA foci in mouse ES cells. DAPI was added to the mounting medium to stain DNA. Scale bars: 10µm. Data are representative of at least three independent experiments with similar results. (B) Boxplot showing 5hmC/R-loops PLA foci per nucleus. Horizontal solid lines represent the median values and whiskers correspond to the 10^{th} and 90^{th} 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.

4.3.4. 5hmC-rich loci are prone to DNA damage

Disruption of R-loop homeostasis is a well-described source of genomic instability[61]. For instance, co-transcriptional R-loops increase conflicts between transcription and replication machineries by creating an additional barrier to fork progression[117,119]. 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 the TTS[144]. 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[145]. The individual distribution profiles of γH2AX were analysed over fixed windows of ±10 kbp around the 5hmC peaks detected in the same cells (Figure 19A). 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 19B). Analysis of γH2AX and 5hmC distribution within active genes revealed a low yet statistically significant Pearson correlation coefficient ($p<0.05$) (Figure 19C). Therefore, we identify 5hmC-rich loci as DNA damage hotspots likely due to their ability to favor R-loop formation.

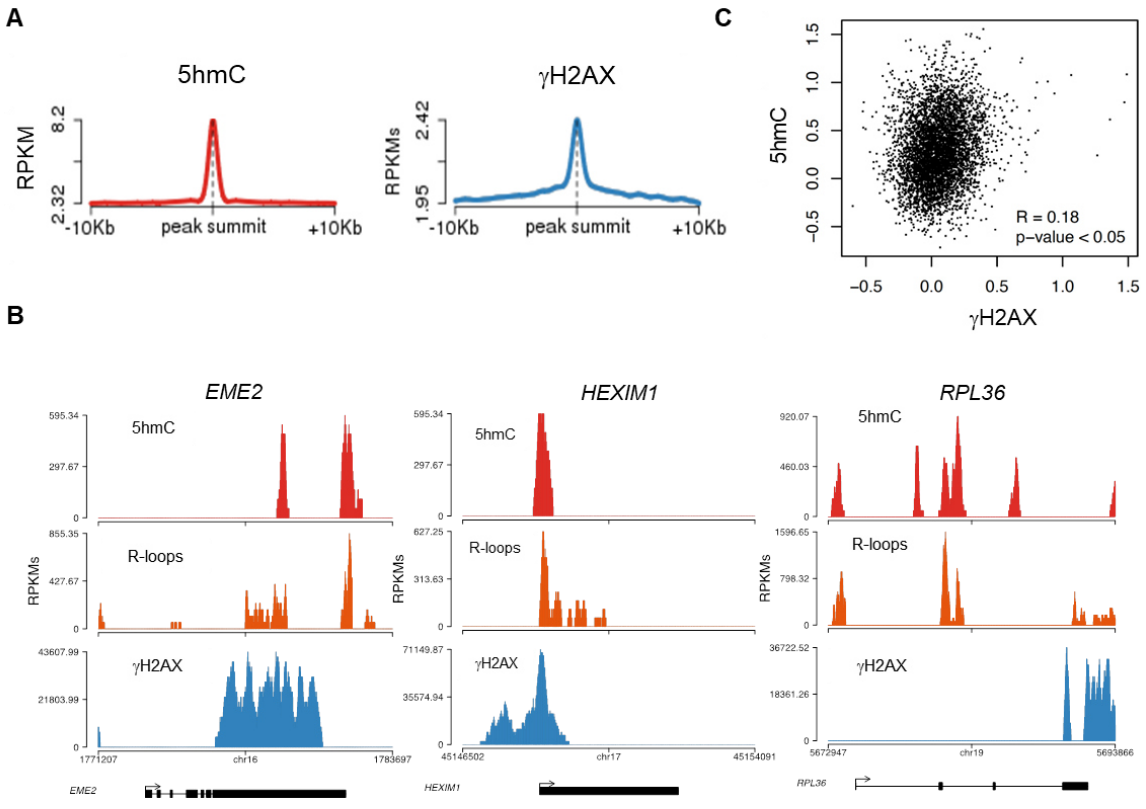


Figure 19: 5hmC-rich loci are genomic hotspots for DNA damage. (A) Metagene profiles of 5hmC and γH2AX probed over fixed windows of ±10 kbp around the 5hmC peaks in expressed genes of HEK293 cells. (B) Individual profiles of 5hmC, R-loops and γH2AX distribution along the *EME2*, *HEXIM1* and *RPL36* genes. Density signals are represented as RPKMs. (C) Pearson correlation coefficient between 5hmC and γH2AX distribution within active genes ($p<0.05$).

4.4. R-loops formed at 5hmC-rich regions impact gene expression in mouse ES cells

To gather insights into the functional impact of R-loops formed at 5hmC-rich DNA regions, we analysed whole-transcriptome (RNA-seq) data of mouse ES cells overexpressing RNase H, a condition resulting in genome-wide loss of R-loops[110]. 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 20A). Pathway analysis revealed that these differentially expressed genes are involved in the mechanistic target of rapamycin (mTOR) (downregulated) and MYC (upregulated) signalling pathways (Figure 20B,C). mTOR and MYC are known to play opposite roles in establishing diapause, the temporary suspension of embryonic development driven by adverse environmental conditions[146], a stage that ES cells mimic when cultured *in vitro*.

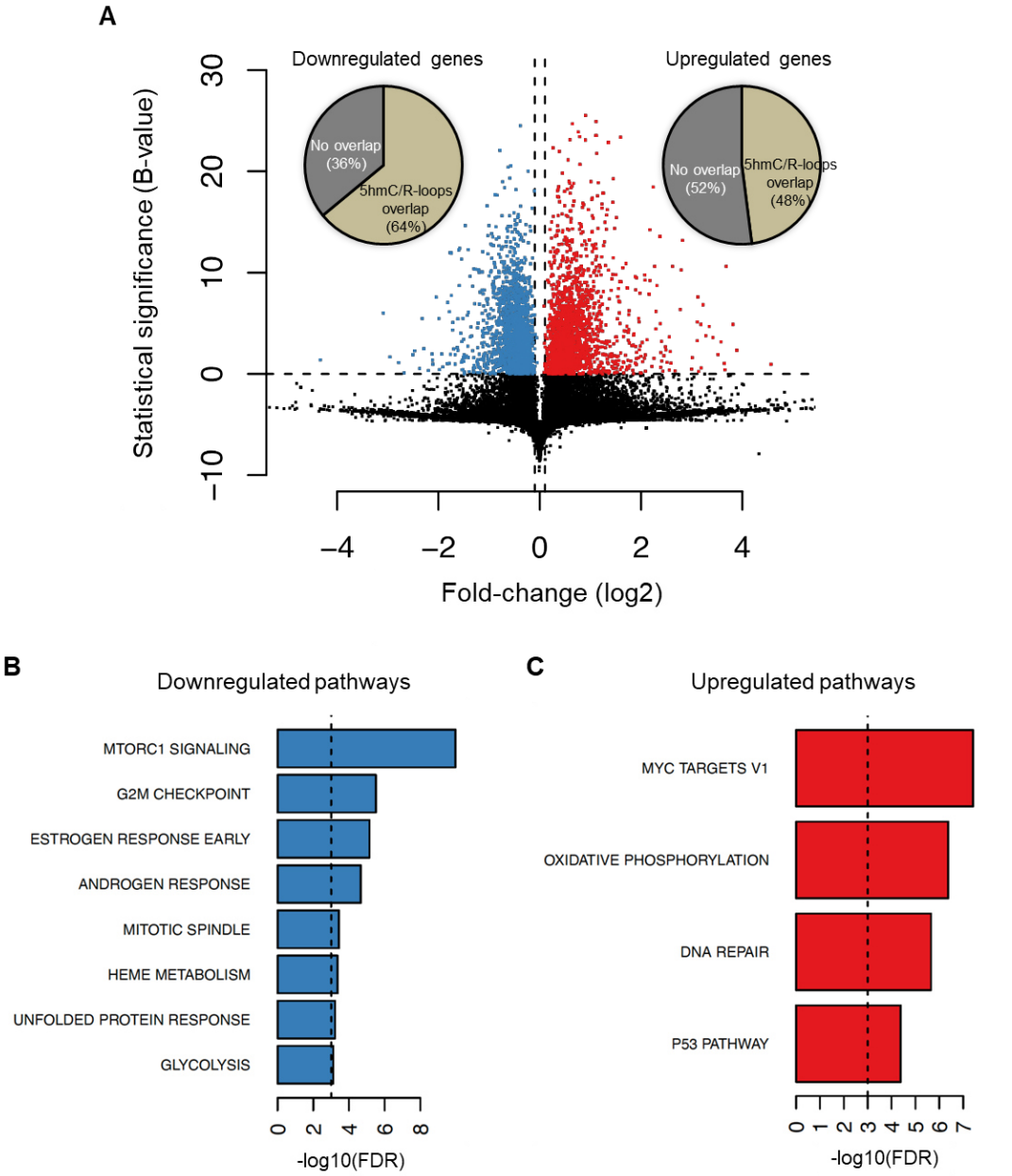


Figure 20: Cellular pathways affected by R-loops formed at 5hmC loci. (A) Volcano plot displaying the differentially expressed genes in mouse ES cells upon RNase H overexpression. Of all downregulated and upregulated genes, 64% and 48% displayed R-loops overlapping with 5hmC, respectively. (B-C) Pathway 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), $p<0.001$.

This raised the hypothesis that RNase H overexpression could interfere with ES cell proliferation. To directly investigate this hypothesis, we overexpressed RNase H in mouse ES cells and analysed cell cycling through measurement of cellular DNA content 24 and 48h later (Figure 21A). No significant changes in cell cycle progression were revealed (Figure

21B). This finding suggests that fine-tuned R-loop formation at specific loci, rather than global changes in R-loop levels, may command the activation of specific gene expression programs in ES cells. In agreement, we observed a significantly decreased expression of genes related to pluripotency (*Oct4*) and germ layer commitment (*Sox17*, *Sox6*, *Dll1*) pathways (Figure 21C). These data support the role of R-loops in the regulation of gene expression in stem cells.

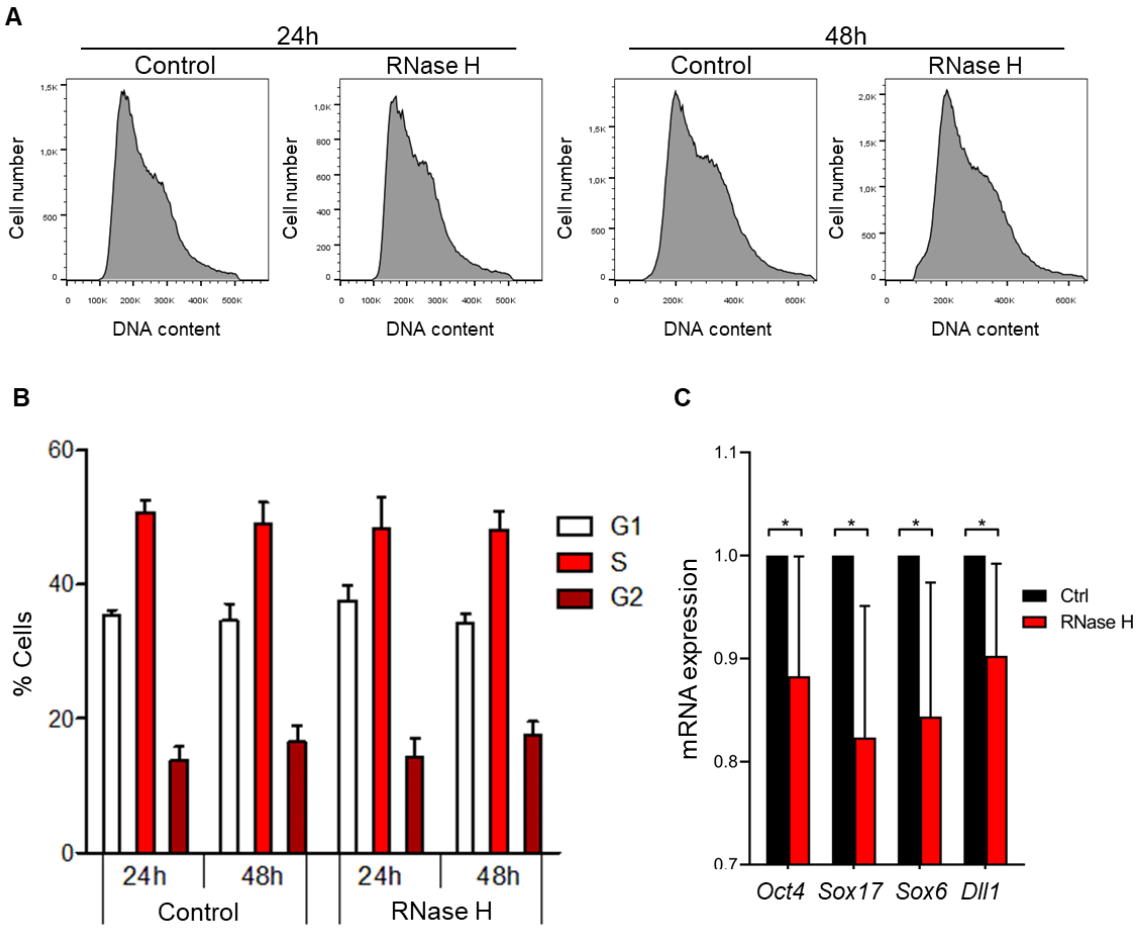


Figure 21: Global R-loop suppression does not impact cell cycle progression, but affects specific genes, in mouse ES cells. (A) Flow cytometry analysis of propidium iodide-treated mouse ES cells with ectopic expression of either GFP (control) or GFP-tagged RNase H for 24 or 48h. Data are representative of five independent experiments. (B) Percentage of control and RNase H-overexpressing mouse ES cells at each cell cycle stage. Means and SDs are from five independent experiments. (C) Transcription levels of pluripotency and germ layer commitment genes in mouse ES cells overexpressing RNase H. *p<0.05, two-tailed Student's t test. Means and SDs are from five independent experiments.

5. Discussion

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

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 provided direct evidence showing that transcription through 5hmC-rich DNA favors R-loop formation. By depleting TET enzymes in mouse ES cells and fibroblasts, we demonstrated that TET activity increases cellular R-loop levels. In agreement, tethering TET enzymes to a specific genomic locus using a CRISPR/Cas9-based system increased the levels of R-loops at the target locus. Notably, the observed variations in R-loop levels did not result from changes in transcription, suggesting that 5hmC directly promotes R-loop formation. Furthermore, our metagene analysis revealed that 5hmC is mostly absent from the TSS. Thus, other chromatin and DNA features (e.g. histone modifications, DNA-supercoiling or G-quadruplex structures[61]) known to induce R-loop formation are likely to operate here. In contrast, the robust overlap between R-loops and 5hmC observed throughout the active genome, peaking at the TTS of active genes, supports a prevailing causal link between these structures.

Mechanistically, 5hmC may impact R-loop formation by either destabilizing the DNA duplex or by altering RNA Pol II elongation rate. Indeed, 5hmC modifies the DNA helix structure by favoring DNA-end breathing motion, diminishes the thermodynamic stability of the DNA duplex and destabilizes GC pairing[50,56]. It also weakens the interaction between DNA and nucleosomal histones[50], which is thought to accelerate RNA Pol II elongation but can also facilitate nascent RNA annealing with the template DNA strand, hence favoring R-loop formation. Future studies can clarify which one of these mechanisms, if not all, contributes to the observed impact of 5hmC on R-loops.

5.2. Interplay between 5hmC and R-loops impinge on multiple cellular processes

As R-loops play diverse physiological roles[61], our findings link TET activity to numerous novel functions, associated with the regulation of gene expression, telomere homeostasis or the maintenance of genome integrity.

5.2.1. Transcription regulation

The observations reported throughout this thesis indicate an effect of the interplay between 5hmC and R-loops in the regulation of gene expression. Our observation that 5hmC and R-loops overlap more robustly at the TTS of active genes supports a model whereby TET enzymes act upstream of R-loop formation during transcription termination. We suggest that TETs deposit 5hmC marks at the TTS, promoting DNA:RNA annealing and the consequent cascade of events that allows efficient termination. In agreement, we observed that *TET1* KO human ES cells exhibit significantly higher readthrough transcripts, a characteristic of transcription termination defects[143]. On the other hand, we found poor overlapping of 5hmC and R-loops at TSSs, where 5hmC marks seem to be underrepresented. However, genome-wide 5hmC mapping across human tissues has revealed that 5hmC peaks at gene promoters[49], so we reason that 5hmC-rich CGI promoters provide a favorable genomic setting for R-loops. Thus, 5hmC-directed R-loop formation at critical gene regulatory regions, such as promoters and gene 3' ends, may govern key steps of the transcription process.

5.2.2. Telomere biology

In the light of our observations, TETs may also impinge on telomere biology. Telomeres are the nucleoprotein complexes found at the ends of linear eukaryotic chromosomes, which can be maintained in proliferating ES and cancer cells by either the activity of telomerase or the alternative lengthening of telomeres (ALT) pathway[147]. ALT telomeres are maintained by mechanisms relying on homologous recombination (HR) between telomeric repeats. R-loops form extensively during transcription of telomeric-repeat-containing RNA (TERRA) and trigger a telomere-specific replication stress, which promotes HR and re-elongation of telomeres by ALT[148,149]. Notably, mouse ES cells depleted of *Tet1* and/or *Tet2* exhibit short telomeres and chromosomal instability, concomitant with reduced telomere recombination[150]. This suggests that telomeric 5hmC might promote HR at telomeres through the establishment of R-loops, which would couple 5hmC/ R-loops functionality to the ALT pathway.

5.2.3. Carcinogenesis

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[151,152]. In addition to altering the expression levels of tumour suppressors or oncogenes[151], 5hmC may indirectly harm genome integrity by promoting R-loop formation and the consequent DNA damage[61], as supported by our observation that 5hmC-rich loci are hotspots for DNA damage genome-wide. Therefore, unsupervised TET-driven changes in the DNA methylation landscape may cause transcription-dependent damaging events that 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[153], suggesting that mis-targeted TET activity may elicit genomic instability, a hallmark of cancer. 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[60]. Thus, we reason that 5hmC-decorated loci may exert both threatening and protective roles over genome integrity, depending on whether TET activity occurs in an unscheduled or controlled fashion, respectively.

5.2.4. ES cell commitment

While the role that TET enzymes play during carcinogenesis is not yet clear, the impact of 5hmC on stem cell differentiation and development has been extensively studied[53]. By driving the developmental DNA methylome reprogramming, TETs carry out numerous functions related to early developmental processes. In this thesis, we disclose a putative new role for R-loops as mediators of 5hmC-driven gene expression programs that determine the self-renewal and differentiation of stem cells. Our gene ontology analysis revealed that R-loops formed at 5hmC-rich regions impact the expression of genes involved in establishing diapause, the temporary suspension of embryonic development driven by adverse environmental conditions[146]. mTOR is a major nutrient sensor that acts as a rheostat during ES cell differentiation, and reductions in mTOR activity trigger diapause[154]. Accordingly, mTOR signalling pathway was significantly downregulated upon global R-loop suppression by RNase H in mouse ES cells. Conversely, MYC targets, which prevent ES cells from entering the state of dormancy that characterizes diapause[155], were amongst the genes more

significantly upregulated upon RNase H overexpression. MYC proteins drive hypertranscription in ES cells, accelerating the gene expression output associated with increased cell proliferation[156]. In agreement with the view that 5hmC-driven R-loop formation impacts functions related to mouse ES cell proliferation, we observed a significant upregulation of genes involved in oxidative phosphorylation (OXPHOS), DNA repair and p53 signalling upon RNase H overexpression. Upregulation of OXPHOS, the main source of energy in most mammalian cells, including ES cells, may fulfil the energetic needs of ES cells resuming proliferation as they exit diapause[157]. Augmented expression of DNA repair and p53 signalling strengthen the genome caretaker and gatekeeper mechanisms that cope with the DNA damage burst observed in highly proliferative cells[158].

This seemingly dichotomous effect of RNase H overexpression in ES cells (i.e. simultaneous decrease of mTOR and increase of MYC signalling) was further corroborated by the lack of significant changes in the proliferation rate of mouse ES cells overexpressing RNase H. Nonetheless, these cells displayed reduced expression levels of genes related to pluripotency (*Oct4*) and germ layer commitment (*Sox17, Sox6, Dll1*) pathways. These findings suggest that programmed 5hmC-driven formation of R-loops at specific genes may constitute a controlled mechanism to direct stem cells' fate. Interestingly, the crosstalk between 5hmC and R-loops has been observed during ES cells differentiation towards neural lineages, a cellular setting deeply characterized by gene reprogramming[159]. Cellular levels of 5hmC and R-loops were inspected throughout ES cells specification, using an *in vitro* adherent monolayer culture system to induce pluripotent mouse ES cells to evolve into neural progenitors competent to initiate neuronal production, which organize into neural tube-like rosettes[160]. Strikingly, data reveals an increase in both 5hmC and R-loops during early neural commitment[159], suggesting that these structures may contribute to the gene output re-shaping that dictates cell specification.

Whether the fine-tuned 5hmC-driven formation of R-loops at specific genes is sufficient to drive their expression and guide stem cells fate, to commit ES cells towards proliferation or diapause establishment, or to dictate cell lineage commitment, 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. Thus, our study sets the ground for further research aimed at investigating the role of the 5hmC/ R-loop axis in gene expression regulation, particularly over ES cells' key pathways.

5.3. Future perspectives

This study raises exciting possibilities about the function of 5hmC-directed R-loop formation on gene expression regulation and in several cellular contexts. Regarding transcription, the findings that TETs absence causes termination impairment led us to suggest a model in which 5hmC assists the formation of R-loops that contribute to efficient termination. However, our data do not provide insights on the impact of 5hmC/ R-loops interplay in transcription initiation. Interestingly, studies have already demonstrated a crosstalk between programmed DNA damage events and the onset of gene transcription[161]. Evidences show an accumulation of DNA damage and γH2AX mark around gene promoters and TSSs upon transcription activation[145,162]. In neural progenitor cells, TSS-proximal DNA DSBs associate with highly transcribed genes[163]. Concomitantly, DSBs genome-wide mapping in cancer cells showed that spontaneous DSBs occur predominantly around the promoter region of actively transcribed genes[164]. Moreover, the observation that promoter site-specific DSBs are required for the transcription of certain genes[165] (such as neuronal early-response genes[166]) reinforces the regulatory role of DNA damage on gene expression. Building on this, since R-loops are well-established sources of genomic instability[61], we envision a mechanism of DNA damage-assisted gene activation in which TET enzymatic activity acts as a new regulatory layer. Following this rationale, we postulate some tentative mechanisms governing early transcription stages.

A master regulatory step of early transcription elongation is RNA Pol II promoter-proximal pause, which occurs through the association of pause-inducing factors: the negative elongation factor (NELF), the DRB-sensitivity-inducing factor (DSIF)[106,107] and TRIM28[167]. Indeed, R-loop enrichment around TSSs often occurs in the context of promoter-proximal pause. Paused RNA Pol II release and productive elongation requires NELF and DSIF phosphorylation by the positive transcription elongation factor b (P-TEFb)[107], as well as TRIM28 phosphorylation, which in the context of promoter-proximal pausing, is mediated by ATM or DNA-PK kinases[167]. Strikingly, R-loops are capable of activating ATM[120]. Hence, we reason that 5hmC deposition over promoter regions favors R-loop formation, which facilitates RNA Pol II release and productive elongation.

Another critical step in early transcription is the determination of transcription direction. Promoter R-loops have already been implicated in the context of divergent transcription, which is characterized by transcription initiation in both directions from the gene promoter, typically in the absence of a reverse-oriented annotated gene[112,168]. Indeed,

studies demonstrate that the majority of unmethylated CpG-rich promoters support divergent transcription, showing an accumulation of transcriptionally engaged polymerase upstream of the annotated gene[169,170]. However, productive elongation usually takes place in the forward direction only, and antisense transcripts (often short and unstable) soon decay [109,171]. Therefore, general mechanisms must be in place to dictate transcription directionality from CGI divergent promoters, in which R-loop formation might be determinant. The evidence herein reported that 5hmC facilitates R-loop formation led us to speculate a mechanism whereby 5hmC-rich promoters provide a permissive milieu for RNA invasion of the DNA double helix. This would create precursor hybrids that elongate preferably towards high GC skew, which is enriched downstream the TSS in CGI-associated promoters[109], thus dictating transcription directionality. Such mechanism would work as a feedback loop: promoter 5hmC endorses persistent DNA:RNA hybrids, which in turn maintain the local chromatin unmethylated[70,91] and contribute to dictate transcription directionality.

Indeed, 5hmC/ R-loops may guide early transcription regulation through an orchestrated effect, whereby transcription elongation in the forward direction is favored by effective promoter-proximal pause surpass. We believe our data paves the way for scrutinizing the so far unexplored coupled effect of such structures in transcription initiation.

5.4. Concluding remarks

The work presented in this thesis unravels a hitherto unappreciated causal link between epigenetic reprogramming by TET enzymes, specifically 5hmC mark, and the formation of DNA:RNA hybrids, with a putative role on gene expression regulation and consequently on physiological processes that rely on extensive gene reprogramming, such as ES cells associated pathways. Thus, our study puts forward promising and exciting lines of research aimed at understanding the role of the 5hmC/ R-loop axis in a plethora of physiological events, and more broadly, at disclosing novel implications of chromatin modifications on DNA secondary structures, which directly impinge on genome functionality.

References

1. Winchester, A. M. Genetics - DNA and the genetic code. *Encyclopædia Britannica, inc.* https://www.britannica.com/science/genetics/DNA-and-the-genetic-code (2020).

2. A Brief History of Genetics: Defining Experiments in Genetics. *Scitable, by Nature Education* https://www.nature.com/scitable/ebooks/a-brief-history-of-genetics-defining-experiments-16570302/contents/ (2014).

3. Gayon, J. De Mendel à l'épigénétique : histoire de la génétique. *Comptes Rendus - Biologies* 339, 225–230 (2016).

4. Deichmann, U. Epigenetics: The origins and evolution of a fashionable topic. *Developmental Biology* 416, 249–254 (2016).

5. Felsenfeld, G. A Brief History of Epigenetics. doi:10.1101/cshperspect.a018200.

6. Rider, C. F. & Carlsten, C. Air pollution and DNA methylation: effects of exposure in humans. *Clinical Epigenetics* (2019) doi:10.1186/s13148-019-0713-2.

7. Egger, G., Liang, G., Aparicio, A. & Jones, P. A. Epigenetics in human disease and prospects for epigenetic therapy. *Nature* vol. 429 457–463 (2004).

8. Wright, M. W. & Bruford, E. A. Naming "junk": Human non-protein coding RNA (ncRNA) gene nomenclature. *Human Genomics* 5, 90–98 (2011).

9. Frías-Lasserre, D. & Villagra, C. A. The Importance of ncRNAs as Epigenetic Mechanisms in Phenotypic Variation and Organic Evolution. *Frontiers in Microbiology* 8, (2017).

10. McGinty, R. K. & Tan, S. Nucleosome Structure and Function. *Chemical Reviews* 115, 2255–2273 (2014).

11. Allfrey, B. V. G., Faulkner, R. & Mirsky, A. E. *Acetylation and Methylation of Histones and Their Possible Role in the Regulation of RNA Synthesis*. *The Rockefeller Institute* vol. 51 (1964).

12. Bannister, A. J. & Kouzarides, T. Regulation of chromatin by histone modifications. *Cell Research* 21, 381–395 (2011).

13. Rossetto, D., Avvakumov, N. & Côté, J. Histone phosphorylation - A chromatin modification involved in diverse nuclear events. *Epigenetics* 7, 1098–1108 (2012).

14. Greer, E. L. & Shi, Y. Histone methylation: a dynamic mark in health, disease and inheritance. *Nature Reviews Genetics* 13, 343–357 (2012).

15. Jin, B. & Robertson, K. D. DNA Methyltransferases (DNMTs), DNA Damage Repair, and Cancer. in *Advances in experimental medicine and biology* vol. 754 3–29 (NIH Public Access, 2013).

16. Ratel, D., Ravanat, J. L., Berger, F. & Wion, D. N6-methyladenine: The other methylated base of DNA. *BioEssays* vol. 28 309–315 (2006).

17. Yu, J., She, Y. & Ji, S. J. m6A Modification in Mammalian Nervous System Development, Functions, Disorders, and Injuries. *Frontiers in Cell and Developmental Biology* 9, 1343 (2021).

18. Greenberg, M. V. C. & Bourc'his, D. The diverse roles of DNA methylation in mammalian development and disease. *Nature Reviews Molecular Cell Biology* 20, 590–607 (2019).

19. Barau, J. *et al.* The DNA methyltransferase DNMT3C protects male germ cells from transposon activity. *Science (1979)* 354, 909–912 (2016).

20. Breiling, A. & Lyko, F. Epigenetic regulatory functions of DNA modifications: 5-methylcytosine and beyond. *Epigenetics and Chromatin* vol. 8 24 (2015).

21. Weber, M. *et al.* Distribution, silencing potential and evolutionary impact of promoter DNA methylation in the human genome. *Nature Genetics* 39, 457–466 (2007).

22. Stadler, M. B. *et al.* DNA-binding factors shape the mouse methylome at distal regulatory regions. *Nature* 480, 490–495 (2011).

23. Yin, Y. *et al.* Impact of cytosine methylation on DNA binding specificities of human transcription factors. *Science (1979)* 356, (2017).

24. Hashimshony, T., Zhang, J., Keshet, I., Bustin, M. & Cedar, H. The role of DNA methylation in setting up chromatin structure during development. *Nature Genetics* 34, 187–192 (2003).

25. Liu, X. *et al.* UHRF1 targets DNMT1 for DNA methylation through cooperative binding of hemi-methylated DNA and methylated H3K9. *Nature Communications* 4, 1513–1563 (2013).

26. Zhao, Q. *et al.* Dissecting the precise role of H3K9 methylation in crosstalk with DNA maintenance methylation in mammals. *Nature Communications* 7, 12464 (2016).

27. Dhayalan, A. *et al.* The Dnmt3a PWWP Domain Reads Histone 3 Lysine 36 Trimethylation and Guides DNA Methylation. *Journal of Biological Chemistry* 285, 26114–26120 (2010).

28. Sun, X. J. *et al.* Identification and characterization of a novel human histone H3 lysine 36-specific methyltransferase. *Journal of Biological Chemistry* 280, 35261–35271 (2005).

29. Shukla, S. *et al.* CTCF-promoted RNA polymerase II pausing links DNA methylation to splicing. *Nature* 479, 74–79 (2011).

30. Maunakea, A. K., Chepelev, I., Cui, K. & Zhao, K. Intragenic DNA methylation modulates alternative splicing by recruiting MeCP2 to promote exon recognition. *Cell Research* 23, 1256–1269 (2013).

31. Pan, Q., Shai, O., Lee, L. J., Frey, B. J. & Blencowe, B. J. Deep surveying of alternative splicing complexity in the human transcriptome by high-throughput sequencing. *Nature Genetics* 40, 1413–1415 (2008).

32. Baralle, F. E. & Giudice, J. Alternative splicing as a regulator of development and tissue identity. *Nature Reviews Molecular Cell Biology* 18, 437–451 (2017).

33. Yearim, A. *et al.* HP1 Is Involved in Regulating the Global Impact of DNA Methylation on Alternative Splicing. *Cell Reports* 10, 1122–1134 (2015).

34. Pastor, W. A., Aravind, L. & Rao, A. TETonic shift: Biological roles of TET proteins in DNA demethylation and transcription. *Nature Reviews Molecular Cell Biology* vol. 14 341–356 (2013).

35. Tahiliani, M. *et al.* Conversion of 5-methylcytosine to 5-hydroxymethylcytosine in mammalian DNA by MLL partner TET1. *Science (1979)* 324, 930–935 (2009).

36. Ito, S. *et al.* Tet proteins can convert 5-methylcytosine to 5-formylcytosine and 5-carboxylcytosine. *Science (1979)* 333, 1300–1303 (2011).

37. He, Y. F. *et al.* Tet-mediated formation of 5-carboxylcytosine and its excision by TDG in mammalian DNA. *Science (1979)* 333, 1303–1307 (2011).

38. Zhang, L. *et al.* Thymine DNA glycosylase specifically recognizes 5-carboxylcytosine-modified DNA. *Nature Chemical Biology* 8, 328–330 (2012).

39. Morgan, H. D., Dean, W., Coker, H. A., Reik, W. & Petersen-Mahrt, S. K. Activation-induced cytidine deaminase deaminates 5-methylcytosine in DNA and is expressed in pluripotent tissues: Implications for epigenetic reprogramming. *Journal of Biological Chemistry* 279, 52353–52360 (2004).

40. Guo, J. U., Su, Y., Zhong, C., Ming, G.-L. & Song, H. Hydroxylation of 5-Methylcytosine by TET1 Promotes Active DNA Demethylation in the Adult Brain. *Cell* 145, 423–434 (2011).

41. Bransteitter, R., Pham, P., Scharfft, M. D. & Goodman, M. F. Activation-induced cytidine deaminase deaminates deoxycytidine on single-stranded DNA but requires the action of RNase. *Proc Natl Acad Sci U S A* 100, 4102–4107 (2003).

42. Nabel, C. S. *et al.* AID/APOBEC deaminases disfavor modified cytosines implicated in DNA demethylation. *Nature Chemical Biology* 8, 751–758 (2012).

43. Ko, M. *et al.* Modulation of TET2 expression and 5-methylcytosine oxidation by the CXXC domain protein IDAX. *Nature* 497, 122–126 (2013).

44. Xu, Y. *et al.* Genome-wide regulation of 5hmC, 5mC, and gene expression by Tet1 hydroxylase in mouse embryonic stem cells. *Mol Cell* 42, 451–64 (2011).

45. Xu, Y. *et al.* Tet3 CXXC domain and dioxygenase activity cooperatively regulate key genes for Xenopus eye and neural development. *Cell* 151, 1200–13 (2012).

46. Arand, J. *et al.* Tet enzymes are essential for early embryogenesis and completion of embryonic genome activation. *EMBO Rep* 23, (2022).

47. An, J. *et al.* Acute loss of TET function results in aggressive myeloid cancer in mice. *Nature Communications* 6, (2015).

48. Tan, L. & Shi, Y. G. Tet family proteins and 5-hydroxymethylcytosine in development and disease. *Development* 139, 1895–1902 (2012).

49. Cui, X. L. *et al.* A human tissue map of 5-hydroxymethylcytosines exhibits tissue specificity through gene and enhancer modulation. *Nature Communications* 11, (2020).

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

51. Hahn, M. A. *et al.* Dynamics of 5-Hydroxymethylcytosine and Chromatin Marks in Mammalian Neurogenesis. *Cell Reports* 3, 291–300 (2013).

52. Yao, B. & Jin, P. Unlocking epigenetic codes in neurogenesis. *Genes and Development* vol. 28 1253–1271 (2014).

53. Ficz, G. *et al.* Dynamic regulation of 5-hydroxymethylcytosine in mouse ES cells and during differentiation. *Nature* 473, 398–404 (2011).

54. Ito, S. *et al.* Role of Tet proteins in 5mC to 5hmC conversion, ES-cell self-renewal and inner cell mass specification. *Nature 2010* 466, 1129–1133 (2010).

55. Sun, W., Zang, L., Shu, Q. & Li, X. From development to diseases: The role of 5hmC in brain. *Genomics* 104, 347–351 (2014).

56. Wanunu, M. *et al.* Discrimination of methylcytosine from hydroxymethylcytosine in DNA molecules. *J Am Chem Soc* 133, 486–492 (2011).

57. Kulaeva, O. I., Gaykalova, D. & Studitsky, V. M. Transcription Through Chromatin by RNA polymerase II: Histone Displacement and Exchange. *Mutation Research - Fundamental and Molecular Mechanisms of Mutagenesis* 618, 116–129 (2007).

58. Levchenko, V., Jackson, B. & Jackson, V. Histone release during transcription: Displacement of the two H2A-H2B dimers in the nucleosome is dependent on different levels of transcription-induced positive stress. *Biochemistry* 44, 5357–5372 (2005).

59. Leavitt, R., Yen, J. & Jia, X.-Y. 5-methylcytosine and 5-hydroxymethylcytosine Exert Opposite Forces on Base Pairing of DNA Double Helix. *Zymo Research Corporation* 6–7 (2015).

60. Kafer, G. R. *et al.* 5-Hydroxymethylcytosine Marks Sites of DNA Damage and Promotes Genome Stability. *Cell Reports* 14, 1283–1292 (2016).

61. García-Muse, T. & Aguilera, A. R Loops: From Physiological to Pathological Roles. *Cell* 179, 604–618 (2019).

62. Costantino, L. & Koshland, D. The Yin and Yang of R-loop biology. *Current Opinion in Cell Biology* 34, 39–45 (2015).

63. Aguilera, A. & García-Muse, T. R Loops: From Transcription Byproducts to Threats to Genome Stability. *Molecular Cell* 46, 115–124 (2012).

64. Xu, B. & Clayton, D. A. RNA-DNA hybrid formation at the human mitochondrial heavy-strand origin ceases at replication start sites: an implication for RNA-DNA hybrids serving as primers. *The EMBO Journal* 15, 3135–3143 (1996).

65. Skourti-Stathaki, K. & Proudfoot, N. J. A double-edged sword: R loops as threats to genome integrity and powerful regulators of gene expression. *Genes and Development* 28, 1384–1396 (2014).

66. Sollier, J. *et al.* Breaking bad: R-loops and genome integrity. *Trends in Cell Biology* 25, 514–522 (2015).

67. Bernecky, C., Herzog, F., Baumeister, W., Plitzko, J. M. & Cramer, P. Structure of transcribing mammalian RNA polymerase II. *Nature* 529, 551–554 (2016).

68. Westover, K. D., Bushnell, D. A. & Kornberg, R. D. Structural Basis of Transcription : Separation of RNA from DNA by RNA Polymerase II. *Science (1979)* 303, 1014–1016 (2004).

69. Chédin, F. Nascent Connections: R-Loops and Chromatin Patterning. *Trends in Genetics* 32, 828–838 (2016).

70. Ginno, P. A., Lim, Y. W., Lott, P. L., Korf, I. & Chédin, F. GC skew at the 5' and 3' ends of human genes links R-loop formation to epigenetic regulation and transcription termination. *Genome Res* 23, 1590–1600 (2013).

71. Maizels, N. & Gray, L. T. The G4 Genome. *PLoS Genetics* 9, e1003468 (2013).

72. Roy, D., Zhang, Z., Lu, Z., Hsieh, C.-L. & Lieber, M. R. Competition between the RNA transcript and the nontemplate DNA strand during R-loop formation in vitro: a nick can serve as a strong R-loop initiation site. *Mol Cell Biol* 30, 146–59 (2010).

73. Bentley, D. L. Coupling mRNA processing with transcription in time and space. *Nature Reviews Genetics* 15, 163–175 (2014).

74. Domínguez-Sánchez, M. S., Barroso, S., Gómez-González, B., Luna, R. & Aguilera, A. Genome Instability and Transcription Elongation Impairment in Human Cells Depleted of THO/TREX. *PLoS Genetics* 7, e1002386 (2011).

75. Huertas, P. & Aguilera, A. Cotranscriptionally Formed DNA:RNA Hybrids Mediate Transcription Elongation Impairment and Transcription-Associated Recombination. *Molecular Cell* 12, 711–721 (2003).

76. Naro, C., Bielli, P., Pagliarini, V. & Sette, C. The interplay between DNA damage response and RNA processing: the unexpected role of splicing factors as gatekeepers of genome stability. *Frontiers in Genetics* 6, (2015).

77. Li, X. & Manley, J. L. Inactivation of the SR Protein Splicing Factor ASF/SF2 Results in Genomic Instability. *Cell* 122, 365–378 (2005).

78. Champoux, J. J. DNA Topoisomerases: Structure, Function, and Mechanism. *Annual Review of Biochemistry* 70, 369–413 (2001).

79. Yang, Y. *et al.* Arginine Methylation Facilitates the Recruitment of TOP3B to Chromatin to Prevent R Loop Accumulation. *Molecular Cell* 53, 484–497 (2014).

80. Tuduri, S. *et al.* Topoisomerase I suppresses genomic instability by preventing interference between replication and transcription. *Nature Cell Biology* 11, 1315–1324 (2009).

81. Shaban, N. M., Harvey, S., Perrino, F. W. & Hollis, T. The structure of the mammalian RNase H2 complex provides insight into RNA·DNA hybrid processing to prevent immune dysfunction. *Journal of Biological Chemistry* 285, 3617–3624 (2010).

82. Cerritelli, S. M. & Crouch, R. J. Ribonuclease H: the enzymes in eukaryotes. *FEBS Journal* 276, 1494–1505 (2009).

83. Zimmer, A. D. & Koshland, D. Differential roles of the RNases H in preventing chromosome instability. *Proceedings of the National Academy of Sciences* 113, 12220–12225 (2016).

84. Sparks, J. L. *et al.* RNase H2-Initiated Ribonucleotide Excision Repair. *Molecular Cell* 47, 980–986 (2012).

85. Cerritelli, S. M. *et al.* Failure to Produce Mitochondrial DNA Results in Embryonic Lethality in Rnaseh1 Null Mice. *Molecular Cell* 11, 807–815 (2003).

86. Ohle, C. *et al.* Transient RNA-DNA Hybrids Are Required for Efficient Double-Strand Break Repair. *Cell* 167, 1001–1013 (2016).

87. Hong, X., Cadwell, G. W. & Kogoma, T. Escherichia coli RecG and RecA proteins in R-loop formation. *The EMBO Journal* 14, 2385–2392 (1995).

88. Boulé, J.-B. & Zakian, V. A. The yeast Pif1p DNA helicase preferentially unwinds RNA-DNA substrates. *Nucleic Acids Research* 35, 5809–5818 (2007).

89. Chakraborty, P. & Grosse, F. Human DHX9 helicase preferentially unwinds RNA-containing displacement loops (R-loops) and G-quadruplexes. *DNA Repair* 10, 654–665 (2011).

90. Chan, Y. A., Hieter, P. & Stirling, P. C. Mechanisms of genome instability induced by RNA processing defects. *Trends in Genetics* 30, 245–253 (2014).

91. Ginno, P. A., Lott, P. L., Christensen, H. C., Korf, I. & Chédin, F. R-Loop Formation Is a Distinctive Characteristic of Unmethylated Human CpG Island Promoters. *Molecular Cell* 45, (2012).

92. Proudfoot, N. J. How RNA polymerase II terminates transcription in higher eukaryotes. *Trends in Biochemical Sciences* 14, 105–110 (1989).

93. Proudfoot, N. J. Transcriptional termination in mammals: Stopping the RNA polymerase II juggernaut. *Science (1979)* 352, 1291- (2016).

94. Epshtein, V., Cardinale, C. J., Ruckenstein, A. E., Borukhov, S. & Nudler, E. An Allosteric Path to Transcription Termination. *Molecular Cell* 28, 991–1001 (2007).

95. Rosonina, E., Kaneko, S. & Manley, J. L. Terminating the transcript: breaking up is hard to do. *Genes & Development* 1050–1056 (2006) doi:10.1101/gad.1431606.

96. Skourti-Stathaki, K., Proudfoot, N. J. & Gromak, N. Human Senataxin Resolves RNA/DNA Hybrids Formed at Transcriptional Pause Sites to Promote Xrn2-Dependent Termination. *Molecular Cell* 42, 794–805 (2011).

97. Kawauchi, J., Mischo, H., Braglia, P., Rondon, A. & Proudfoot, N. J. Budding yeast RNA polymerases I and II employ parallel mechanisms of transcriptional termination. *Genes & Development* 1082–1092 (2008) doi:10.1101/gad.463408.

98. Mischo, H. E. *et al.* Yeast Sen1 helicase protects the genome from transcription-associated instability. *Mol Cell* 41, 21–32 (2011).

99. Bonnet, A. *et al.* Introns Protect Eukaryotic Genomes from Transcription-Associated Genetic Instability. *Molecular Cell* 67, 608–621 (2017).

100. Niu, D.-K. Protecting exons from deleterious R-loops: a potential advantage of having introns. *Biol Direct* 2, (2007).

101. Sridhara, S. C. *et al.* Transcription Dynamics Prevent RNA-Mediated Genomic Instability through SRPK2-Dependent DDX23 Phosphorylation. *Cell Reports* 18, 334–343 (2017).

102. Bernstein, E. & Allis, C. D. RNA meets chromatin. *Genes & Development* 1635–1655 (2005) doi:10.1101/gad.1324305.GENES.

103. Nakama, M., Kawakami, K., Kajitani, T., Urano, T. & Murakami, Y. DNA-RNA hybrid formation mediates RNAi-directed heterochromatin formation. *Genes to Cells* 17, 218–233 (2012).

104. Skourti-Stathaki, K., Kamieniarz-Gdula, K. & Proudfoot, N. J. R-loops induce repressive chromatin marks over mammalian gene terminators. *Nature* 516, 436–439 (2014).

105. Castellano-Pozo, M. *et al.* R loops are linked to histone H3 S10 phosphorylation and chromatin condensation. *Molecular Cell* 52, 583–590 (2013).

106. Adelman, K. & Lis, J. T. Promoter-proximal pausing of RNA polymerase II: emerging roles in metazoans. *Nature Reviews Genetics* 13, 720–731 (2012).

107. Jonkers, I. & Lis, J. T. Getting up to speed with transcription elongation by RNA polymerase II. *Nature Reviews Molecular Cell Biology* 16, 167–177 (2015).

108. Shivji, M. K. K., Renaudin, X., Williams, Ç. H. & Venkitaraman, A. R. BRCA2 Regulates Transcription Elongation by RNA Polymerase II to Prevent R-Loop Accumulation. *Cell Reports* 22, 1031–1039 (2018).

109. Kellner, W. A., Bell, J. S. K. & Vertino, P. M. GC skew defines distinct RNA polymerase pause sites in CpG island promoters. *Genome Research* 25, 1600–1609 (2015).

110. Chen, P. B., Chen, H. v, Acharya, D., Rando, O. J. & Fazzio, T. G. R loops regulate promoter-proximal chromatin architecture and cellular differentiation. *Nat Struct Mol Biol* 22, 999–1007 (2015).

111. Grunseich, C. *et al.* Senataxin Mutation Reveals How R-Loops Promote Transcription by Blocking DNA Methylation at Gene Promoters. *Molecular Cell* 69, 426–437 (2018).

112. Boque-Sastre, R. *et al.* Head-to-head antisense transcription and R-loop formation promotes transcriptional activation. *Proc Natl Acad Sci U S A* 112, 5785–5790 (2015).

113. Arab, K. *et al.* GADD45A binds R-loops and recruits TET1 to CpG island promoters. *Nature Genetics* 51, 217–223 (2019).

114. Kuzminov, A. Single-strand interruptions in replicating chromosomes cause double-strand breaks. *Proc Natl Acad Sci U S A* 98, 8241–8246 (2001).

115. Aguilera, A. The connection between transcription and genomic instability. *EMBO Journal* 21, 195–201 (2002).

116. Dunn, K. & Griffith, J. D. The presence of RNA in a double helix inhibits its interaction with histone protein. *Nucleic Acids Research* 8, 555–566 (1980).

117. Hamperl, S., Bocek, M. J., Saldivar, J. C., Swigut, T. & Cimprich, K. A. Transcription-Replication Conflict Orientation Modulates R-Loop Levels and Activates Distinct DNA Damage Responses. *Cell* 170, 774–786 (2017).

118. Gaillard, H., García-Muse, T. & Aguilera, A. Replication stress and cancer. *Nature Reviews Cancer* 15, 276–280 (2015).

119. Helmrich, A., Ballarino, M., Nudler, E. & Tora, L. Transcription-replication encounters, consequences and genomic instability. *Nature Structural and Molecular Biology* 20, 412–418 (2013).

120. Tresini, M. *et al.* The core spliceosome as target and effector of non-canonical ATM signalling. *Nature* 523, 53–58 (2015).

121. Smith, A. G. & Hooper, M. L. Buffalo rat liver cells produce a diffusible activity which inhibits the differentiation of murine embryonal carcinoma and embryonic stem cells. *Developmental Biology* 121, 1–9 (1987).

122. Langmead, B., Trapnell, C., Pop, M. & Salzberg, S. L. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biology* 10, (2009).

123. Zhang, Y. *et al.* Model-based analysis of ChIP-Seq (MACS). *Genome Biology* 9, (2008).

124. Quinlan, A. R. & Hall, I. M. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics* 26, 841–842 (2010).

125. Althammer, S., González-Vallinas, J., Ballaré, C., Beato, M. & Eyras, E. Pyicos: a versatile toolkit for the analysis of high-throughput sequencing data. *Bioinformatics* 27, 3333–3340 (2011).

126. R Core Team (2018). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Austria.

127. Li, H. *et al.* The Sequence Alignment/Map format and SAMtools. *Bioinformatics* 25, 2078–2079 (2009).

128. Bray, N. L., Pimentel, H., Melsted, P. & Pachter, L. Near-optimal probabilistic RNA-seq quantification. *Nature Biotechnology* 34, 525–527 (2016).

129. Robinson, M. D., McCarthy, D. J. & Smyth, G. K. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics* 26, 139–140 (2010).

130. Ritchie, M. E. *et al.* limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic Acids Research* 43, e47 (2015).

131. Liberzon, A. *et al.* The Molecular Signatures Database Hallmark Gene Set Collection. *Cell Systems* 1, 417–425 (2015).

132. Dobin, A. *et al.* STAR: Ultrafast universal RNA-seq aligner. *Bioinformatics* 29, 15–21 (2013).

133. Roth, S. J., Heinz, S. & Benner, C. ARTDeco: Automatic readthrough transcription detection. *BMC Bioinformatics* 21, 1–22 (2020).

134. Ramírez, F. *et al.* deepTools2: a next generation web server for deep-sequencing data analysis. *Nucleic Acids Res* 44, W160–W165 (2016).

135. Carrasco-Salas, Y. *et al.* The extruded non-template strand determines the architecture of R-loops. *Nucleic Acids Research* 47, 6783–6795 (2019).

136. Klinov, D. V *et al.* High resolution mapping DNAs by R-loop atomic force microscopy. *Nucleic Acids Research* 26, 4603–4610 (1998).

137. Liu, X. S. *et al.* Editing DNA Methylation in the Mammalian Genome. *Cell* 167, 233–247 (2016).

138. Sanz, L. A. *et al.* Prevalent, Dynamic, and Conserved R-Loop Structures Associate with Specific Epigenomic Signatures in Mammals. *Molecular Cell* 63, 167–178 (2016).

139. Chen, L. *et al.* R-ChIP Using Inactive RNase H Reveals Dynamic Coupling of R-loops with Transcriptional Pausing at Gene Promoters. *Molecular Cell* 68, 745-757.e5 (2017).

140. Matarese, F., Carrillo-De Santa Pau, E. & Stunnenberg, H. G. 5-Hydroxymethylcytosine: a new kid on the epigenetic block? *Molecular Systems Biology* 7, (2011).

141. Jin, C. *et al.* TET1 is a maintenance DNA demethylase that prevents methylation spreading in differentiated cells. *Nucleic Acids Research* 42, 6956–6971 (2014).

142. Nadel, J. *et al.* RNA:DNA hybrids in the human genome have distinctive nucleotide characteristics, chromatin composition, and transcriptional relationships. *Epigenetics & Chromatin* 8, (2015).

143. Nojima, T. & Proudfoot, N. J. Mechanisms of lncRNA biogenesis as revealed by nascent transcriptomics. *Nature Reviews Molecular Cell Biology* 0123456789, (2022).

144. Hatchi, E. *et al.* BRCA1 recruitment to transcriptional pause sites is required for R-loop-driven DNA damage repair. *Molecular Cell* 57, 636–647 (2015).

145. Bunch, H. *et al.* Transcriptional elongation requires DNA break-induced signalling. *Nature Communications* 6, (2015).

146. Fenelon, J. C., Banerjee, A. & Murphy, B. D. Embryonic diapause: development on hold. *International Journal of Developmental Biology* 58, 163–174 (2014).

147. Claude, E. & Decottignies, A. Telomere maintenance mechanisms in cancer: telomerase, ALT or lack thereof. *Current Opinion in Genetics & Development* 60, 1–8 (2020).

148. Arora, R. *et al.* RNaseH1 regulates TERRA-telomeric DNA hybrids and telomere maintenance in ALT tumour cells. *Nature Communications* 5, 1–11 (2014).

149. Domingues-Silva, B., Silva, B. & Azzalin, C. M. ALTernative Functions for Human FANCM at Telomeres. *Frontiers in Molecular Biosciences* 6, (2019).

150. Yang, J. *et al.* Tet Enzymes Regulate Telomere Maintenance and Chromosomal Stability of Mouse ESCs. *Cell Reports* 15, 1809–1821 (2016).

151. Bray, J. K., Dawlaty, M. M., Verma, A. & Maitra, A. Roles and Regulations of TET Enzymes in Solid Tumors. *Trends in Cancer* (2021) doi:10.1016/j.trecan.2020.12.011.

152. Good, C. R. *et al.* TET1-Mediated Hypomethylation Activates Oncogenic Signaling in Triple-Negative Breast Cancer. *Cancer Research* 78, 4126–4137 (2018).

153. Good, C. R. *et al.* A novel isoform of TET1 that lacks a CXXC domain is overexpressed in cancer. *Nucleic Acids Research* 45, 8269–8281 (2017).

154. Bulut-Karslioglu, A. *et al.* Inhibition of mTOR induces a paused pluripotent state. *Nature* 540, 119–123 (2016).

155. Scognamiglio, R. *et al.* Myc Depletion Induces a Pluripotent Dormant State Mimicking Diapause. *Cell* 164, 668–680 (2016).

156. Percharde, M., Bulut-Karslioglu, A. & Ramalho-Santos, M. Hypertranscription in Development, Stem Cells, and Regeneration. *Developmental Cell* 40, 9–21 (2017).

157. Mathieu, J. & Ruohola-Baker, H. Metabolic remodeling during the loss and acquisition of pluripotency. *Development* 144, 541–551 (2017).

158. Williams, A. B. & Schumacher, B. p53 in the DNA-Damage-Repair Process. *Cold Spring Harbor Perspectives in Medicine* 1–15 (2016).

159. Sabino, J. C. Epigenetic drivers of genomic instability during neural differentiation. (2016).

160. Abranches, E. *et al.* Neural differentiation of embryonic stem cells in vitro: A road map to neurogenesis in the embryo. *PLoS ONE* (2009) doi:10.1371/journal.pone.0006286.

161. Fong, Y. W., Cattoglio, C. & Tjian, R. The Intertwined Roles of Transcription and Repair Proteins. *Molecular Cell* 52, 291–302 (2013).

162. Lensing, S. V *et al.* DSBCapture: in situ capture and sequencing of DNA breaks. *Nature Methods* 13, 1–6 (2016).

163. Schwer, B. *et al.* Transcription-associated processes cause DNA double-strand breaks and translocations in neural stem/progenitor cells. *Proceedings of the National Academy of Sciences* 113, 2258–2263 (2016).

164. Yang, F., Kemp, C. J. & Henikoff, S. Anthracyclines induce double-strand DNA breaks at active gene promoters. *Mutation Research - Fundamental and Molecular Mechanisms of Mutagenesis* 773, 9–15 (2015).

165. Ju, B.-J. *et al.* A Topoisomerase IIβ -Mediated dsDNA Break Required for Regulated Transcription. *Science (1979)* 312, 1798–1802 (2006).

166. Madabhushi, R. *et al.* Activity-Induced DNA Breaks Govern the Expression of Neuronal Early-Response Genes. *CELL* 161, 1592–1605 (2015).

167. Bunch, H. *et al.* TRIM28 regulates RNA polymerase II promoter-proximal pausing and pause release. *Nat Struct Mol Biol* 21, 876–883 (2014).

168. Core, L. J., Waterfall, J. J. & Lis, J. T. Nascent RNA Sequencing Reveals Widespread Pausing and Divergent Initiation at Human Promoters. *Science (1979)* 322, 0–4 (2008).

169. Wu, X. & Sharp, P. A. Divergent Transcription: A Driving Force for New Gene Origination? *Cell* 155, 990–996 (2013).

170. Core, L. J. *et al.* Defining the Status of RNA Polymerase at Promoters. *Cell Reports* 2, 1025–1035 (2012).

171. Duttke, S. H. C. *et al.* Human Promoters Are Intrinsically Directional. *Molecular Cell* 57, 674–684 (2015).

172. Baedke, J. The epigenetic landscape in the course of time: Conrad Hal Waddington's methodological impact on the life sciences. *Studies in History and Philosophy of Science Part C :Studies in History and Philosophy of Biological and Biomedical Sciences* 44, 756–773 (2013).

6. Annexes

Supplementary Table 1: shRNA sequence.

Gene knockdown	shRNA sequence
Tet3	tgctgttgacagtgagcgcgcagtgtgtattcctaccatttagtgaagccacagatgtaaatggtaggaatacacactgcttgcctactgcctcgga

Supplementary Table 2: Oligonucleotide sequences.

Primers	Sequence
M13 FOR long	GTTTTCCCAGTCACGACGTTGT
M13 REV long	AACAGCTATGACCATGATTACGCCA
[γ-32P] M13 FOR long	GTTTTCCCAGTCACGACGTTGT
Tet1 Transcript FW	GAAGGTATCCCTCGCCTGAT
Tet1 Transcript RV	CCACGAACAGCCAAAGGAGA
Tet2 Transcript FW	GTCAACATGCCAGGAGGATTC
Tet2 Transcript RV	TTGCGGGTGAGCCTCAGATG
Tet3 Transcript FW	ACACCCTCTACCAGGAGCTT
Tet3 Transcript RV	GCAGCCGTTGAAGTACATGC
Smad9_DRIP_RNA FW	CACAGCGAGTACAACCCTCA
Smad9_DRIP_RNA RV	ATGGAGACTGCGGAAACACA
Diexf_DRIP_RNA FW	ATGCGATAGCTCTTGGGAGG
Diexf_DRIP_RNA RV	TTCAACCCGCCCTTCCATTT
Tom1l1_DRIP FW	CACATGGGTCTTACAGACAG
Tom1l1_DRIP RV	GAGTTTGGGATGCTGGTGAT
Tom1l1_RNA FW	TTCTGTTCTGGGTCTCCAGC
Tom1l1_RNA RV	ATGTGCGTGCAGAACTGTGG
Slc8a1_DRIP_RNA FW	GTCCATTGCTGCCATCTACCA
Slc8a1_DRIP_RNA RV	GAAGATGTGAGGAGCTTGGCA
Actb_DRIP_RNA FW	GAACCGCTCGTTGCCAATAG
Actb_DRIP_RNA RV	CACCACAGCTGAGAGGGAAA
Hprt1_DRIP_RNA FW	GTCATGAAGGAGATGGGAGG
Hprt1_DRIP_RNA RV	ATGTAATCCAGCAGGTCAGC
Palm2_DRIP FW	CAGCTTTATCCTGGGCGTGA

Palm2_DRIP RV	TCATCCCGCATCCTATGCAC
Palm2_RNA FW	ACAATGGCCTCCTCGCTGAT
Palm2_RNA RV	ATCACAGCAGTTGGCCTCCA
Gpr180_DRIP FW	ACCTTGATACACGCGCTCTT
Gpr180_DRIP RV	TGCTGACCTTGACAATCGAC
Gpr180_RNA FW	AGACGGAGCACAACCTCACA
Gpr180_RNA RV	CGGGTTGAGGAGCACCATTT
Srpk1_DRIP FW	TGCTCACAACGTCTTACCCA
Srpk1_DRIP RV	CCTCCTGAGACCAAGATTCT
Srpk1_RNA FW	AGTCATTGGGGTCTTCCTGC
Srpk1_RNA RV	TGAAACTCAGCACCGAGGCT
Hip1r_DRIP FW	TGCATGACTATCAGCGGTAC
Hip1r_DRIP RV	TTATGGAGCTGTCGGCCAAT
Hip1r_RNA FW	AGGGAGCCTTTACCTTCTGG
Hip1r_RNA RV	GAGGACCTTGTGAAGGACGT
B2m_DRIP_RNA FW	ACGTAACACAGTTCCACCCG
B2m_DRIP_RNA RV	TCAGTCTCAGTGGGGGTGAA
APOE Pair A FW	GCTGCGTTGCTGGTCACATT
APOE Pair A RV	CAGGAGGTTGAGGTGAGGAT
APOE Pair B_1 FW	GCCCGAGCTGCGCCAG
APOE Pair B_1 RV	ACAGTGTCTGCACCCAGC
APOE Pair B_2_mRNA FW	GGCAGAGCGGCCAGCG
APOE Pair B_2_mRNA RV	CTCCTCCTGCACCTGCTC
APOE Pair C_1 FW	GCCTACAAATCGGAACTGGA
APOE Pair C_1 RV	CAGCTCCTCGGTGCTCTG
APOE Pair C_2 FW	CCGTTCCTTCTCTCCCTCTT
APOE Pair C_2 RV	TCCAGTTCCGATTTGTAGGC
APOE Pair D FW	TGAAGGAGCAGGTGGCGGA
APOE Pair D RV	CTGGCGCTGCATGTCTTCCA
Oct4 FW	GAAGCCGACAACAATGAGAACC
Oct4 RV	CTCCAGACTCCACCTCACACG
Sox17 FW	ACAACGCAGAGCTAAGCAAGAT
Sox17 RV	GTACTTGTAGTTGGGGTGGTCCT

Sox6 FW	TCAACCTGCCAAACAAAAGC
Sox6 RV	GCTGGATCTGTTCTCGCATC
Dll1 FW	GCAGGACCTTCTTTCGCGTAT
Dll1 RV	AAGGGGAATCGGATGGGGTT
U6 snRNA FW	GCTTCGGCAGCACATATACTA
U6 snRNA RV	AAATATGGAACGCTTCACGA
Gapdh FW	AACTTTGGCATTGTGGAAGG
Gapdh RV	ACACATTGGGGGTAGGAACA

Supplementary Table 3: Antibodies used in this study.

Product	**Concentrations**	**Company/ Cat. No.**	**Notes**
S9.6	5ug/ IP; 1:1000 (DB)	Millipore; MABE1095	Anti-DNA:RNA hybrid antibody used to detect R-loops
dsDNA	1:1000 (DB)	Santa Cruz; sc-58749	Anti-dsDNA specific antibody (HYB331-01)
5hmC	5ug/ IP; 1:1000 (DB)	Active Motif; 39791	5-hydroxymethylcytosine antibody
5mC	5ug/ IP; 1:500 (DB)	Active Motif; 61255	5-methylcytosine antibody

Supplementary Table 4: g-blocks sequences.

g-blocks
β-actin P1
CTGACAACCGGTGTTTTCCCAGTCACGACGTTGTTAATACGACTCACTATAG
GGTTACCCAGAGTGCAGGTGTGTGGAGATCCCTCCTGCCTTGACATTGAGCA
GCCTTAGAGGGTGGGGGAGGCTCAGGGGTCAGGTCTCTGTTCCTGCTTATTG
GGGAGTTCCTGGCCTGGCCCTTCTATGTCTCCCCAGGTACCCCAGTTTTTCTG
GGTTCACCCAGAGTGCAGATGCTTGAGGAGGTGGGAAGGGACTATTTGGGG
GTGTCTGGCTCAGGTGCCATGCCTCACTGGGGCTGGTTGGCACCTGCATTTC
CTGGGAGTGGGGCTGTCTCAGGGTAGCTGGGCACGGTGTTCCCTTGAGTGGG

GGTGTAGTGGGTGTTCCTAGCTGCCACGCCTTTGCCTTCACCTATGGGATCGT GGCTGTCAGCCTTGAGGGTCAGCCTGGCCCAGGCTCCTGGCGTAATCATGGT CATAGCTGTTTGTACACTGACA
β-actin P2
CTGACAACCGGTGTTTTCCCAGTCACGACGTTGTTAATACGACTCACTATAG GGGGGACTATTTGGGGGTGTCTGGCTCAGGTGCCATGCCTCACTGGGGCTGG TTGGCACCTGCATTTCCTGGGAGTGGGGCTGTCTCAGGGTAGCTGGGCACGG TGTTCCCTTGAGTGGGGGTGTAGTGGGTGTTCCTAGCTGCCACGCCTTTGCCT TCACCTATGGGATCGTGGCTGTCAGCCTTGAGGGTCAGCCTGGCCCAGGCTC CCATAGGCTTAGGAGAGGCCGCAATTCCTACCTGTTCATCCAGACAGAGGGG GACCTGGAATCAAAGTCAAGTTGGGGTAGGGGGTCCATGGGGCCATATCTG GCCTGCAGACAGCTCTGGTTAGCTATGGGCTGAGGTCTGGATTCTGCCTTGT GACTGGAGACTGGGCGCCATCCCGTGGCCTCTGAGGGCTGGCGTAATCATGG TCATAGCTGTTTGTACACTGACA
APOE
CTGACAACCGGTGGTTTTCCCAGTCACGACGTTGTAATACGACTCACTATAG GGCCGGTGAGAAGCGCAGTCGGGGGCACGGGGATGAGCTCAGGGGCCTCTA GAAAGAGCTGGGACCCTGGGAACCCCTGGCCTCCAGGTAGTCTCAGGAGAG CTACTCGGGGTCGGGCTTGGGGAGAGGAGGAGCGGGGGTGAGGCAAGCAGC AGGGGACTGGACCTGGGAAGGGCTGGGCAGCAGAGACGACCCGACCCGCTA GAAGGTGGGGTGGGGAGAGCAGCTGGACTGGGATGTAAGCCATAGCAGGAC TCCACGAGTTGTCACTATCATTTATCGAGCACCTACTGGGTGTCCCCAGTGTC CTCAGATCTCCATAACTGGGGAGCCAGGGGCAGCGACACGGTAGCTAGCCG TCGATTGGAGAACTTTAAAATGAGGACTGAATTAGCTCATAAATGGCGTAAT CATGGTCATAGCTGTTTGTACACTGACA

Supplementary Table 5: S9.6 EMSA oligonucleotides.

Single-stranded oligonucleotide	Sequence
ssDNA_S9.6 EMSA	GCTGTCAGAC
ssRNA_S9.6 EMSA	GUCUGACAGC

RESEARCH ARTICLE

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

João C Sabino[1], Madalena R de Almeida[1], Patrícia L Abreu[1], Ana M Ferreira[1], Paulo Caldas[2,3], Marco M Domingues[1], Nuno C Santos[1], Claus M Azzalin[1], Ana Rita Grosso[2,3], Sérgio Fernandes de Almeida[1]*

[1]Instituto de Medicina Molecular João Lobo Antunes, Faculdade de Medicina, Universidade de Lisboa, Lisbon, Portugal; [2]Associate laboratory i4HB – Institute for Health and Bioeconomy, NOVA School of Science and Technology, Universidade Nova de Lisboa, Caparica, Portugal; [3]UCIBIO-REQUIMTE, Applied Molecular Biosciences Unit, Department of Life Sciences, NOVA School of Science and Technology, Universidade Nova de Lisboa, Lisbon, Portugal

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 the 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 shows 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.

***For correspondence:** sergioalmeida@fm.ul.pt

Competing interest: The authors declare that no competing interests exist.

Funding: See page 17

Received: 16 April 2021
Preprinted: 27 April 2021
Accepted: 21 February 2022
Published: 22 February 2022

Reviewing Editor: Andrés Aguilera, CABIMER, Universidad de Sevilla, Spain

Editor's evaluation

The study shows a correlation between 5hmC and R loops in mES cells and human HEK293 cells depleted of the TET enzymes Tet1, Tet2 and Tet3 that convert 5mC into 5hmC. The data presented are clearly of significant interest in providing new insight into the potential role of 5hmc DNA in specific transcriptional processes. This implies that 5hmC containing DNA has specific epigenetic features beyond a simple intermediate in interconversion between repressive 5mC and active C DNA.

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 (***García-Muse and Aguilera, 2019***). However, nonscheduled or persistent R-loops constitute an important source of DNA damage, namely, DNA double-strand

breaks (DSBs) (**García-Muse and Aguilera, 2019**). 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 **García-Muse and Aguilera, 2019**). 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 (**García-Muse and Aguilera, 2019**). 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 (**Sridhara et al., 2017**). 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 (**Bonnet et al., 2017**). 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) favoring R-loop formation (**Ginno et al., 2013**). Moreover, chromatin and DNA features such as histone modifications, DNA-supercoiling, and G-quadruplex structures also affect R-loop establishment (**García-Muse and Aguilera, 2019**). 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 (**Chen et al., 2015**). 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 (**Skourti-Stathaki and Proudfoot, 2014**; **Chédin, 2016**; **Skourti-Stathaki et al., 2014**).

Besides affecting histone modifications, R-loops also act as barriers against DNA methylation spreading into active genes (**Ginno et al., 2013**; **Ginno et al., 2012**). 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) (**Karpf, 2013**). The activity of DNA methyltransferase (DNMT) enzymes makes 5mC widespread across the mammalian genome where it plays major roles in imprinting, retrotransposon silencing, and gene expression (**Greenberg and Bourc'his, 2019**). 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 (**Greenberg and Bourc'his, 2019**; **Weber et al., 2007**). R-loops positioned near promoters of active genes maintain CGIs in an unmethylated state (**Ginno et al., 2012**), likely by reducing the affinity of DNMT1 binding to DNA (**Grunseich et al., 2018**), or recruiting ten-eleven translocation (TET) methylcytosine dioxygenases (**Arab et al., 2019**).

The TET enzyme family members share the ability to oxidize 5mC to 5-hydroxymethylcytosine (5hmC) (**Pastor et al., 2013**; **Tahiliani et al., 2009**). 5hmC is a relatively rare DNA modification found across the genome much less frequently than 5mC (**Mendonca et al., 2014**). 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 (**Wu et al., 2011**). The levels of 5hmC are enriched at active promoter regions, as observed upon activation of neuronal function-related genes in neural progenitors and neurons (**Pastor et al., 2013**; **Hahn et al., 2013**). 5hmC has the potential to modify the DNA helix structure by favoring DNA-end breathing motion, a dynamic feature of the protein–DNA complexes thought to control DNA accessibility (**Mendonca et al., 2014**). 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 (**Mendonca et al., 2014**). While 5mC increases the melting temperature, 5hmC reduces the amount of energy needed to separate the two strands of the DNA duplex (**Leavitt et al., 2015**; **Wanunu et al., 2011**). 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 (**Wanunu et al., 2011**). The presence of 5hmC destabilizes GC pairing by alleviating steric constraints through an increase in molecular polarity (**Wanunu et al., 2011**).

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 favor R-loop formation. Here, we show that 5hmC promotes R-loop formation during in vitro transcription of DNA templates. 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 (γH2AX), 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 favors 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 400 bp sequence containing a genomic region prone to form R-loops in vivo (***Sridhara et al., 2017***; ***Skourti-Stathaki et al., 2014***). 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 favors 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 in vitro transcription reactions (***Figure 1E***). R-loops were identified as previously described (***Carrasco-Salas et al., 2019***; ***Klinov et al., 1998***). 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 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 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 in vivo.

TET enzymatic activity impacts endogenous R-loop levels

To test whether the 5hmC DNA modification induces R-loop formation 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

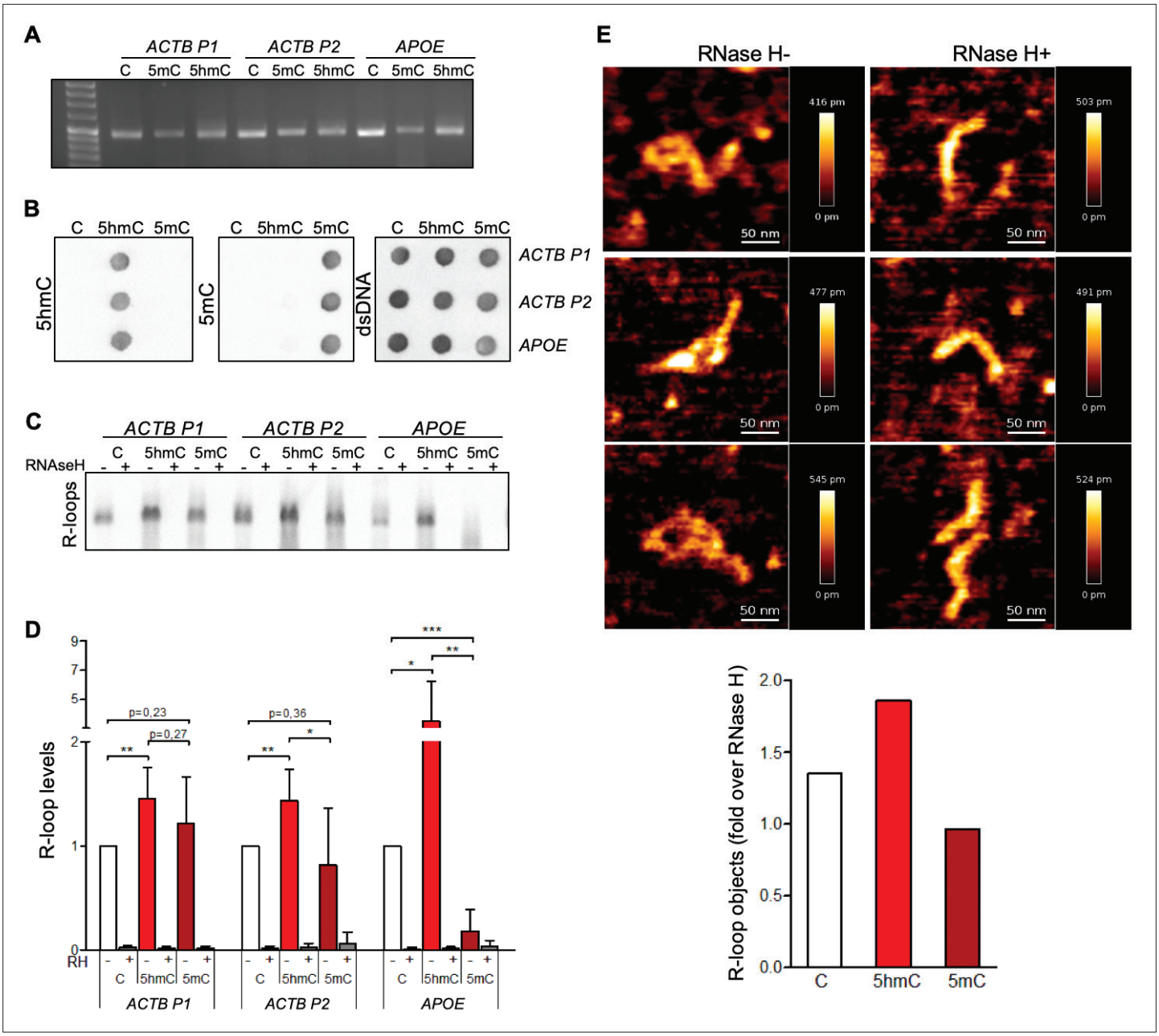


Figure 1. 5-Hydroxymethylcytosine (5hmC) favors co-transcriptional R-loop formation. (**A**) Native or modified deoxyribonucleotides (dCTPs) were incorporated upon PCR amplification into DNA fragments with sequences from the transcription termination region of *ACTB* (*ACTB* P1 and *ACTB* P2) or *APOE*. (**B**) Incorporation of dCTP variants confirmed by immunoblotting using specific antibodies against 5-methylcytosine (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 (AFM). 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.

The online version of this article includes the following figure supplement(s) for figure 1:

Figure supplement 1. Cytosine modifications do not affect the detection of DNA:RNA hybrids by the S9.6 Ab.

Figure supplement 2. In vitro transcription levels of different DNA templates.

Tet1 or *Tet2* depletion (***Figure 2A***, ***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***, ***Figure 2—figure supplement 1B***). No significant changes were observed in 5mC levels (***Figure 2A***, ***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***, ***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 and 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 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 analyzed genes in mES cells and 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 the absence of changes in transcription levels.

Next, we employed a modified CRISPR-based system to target TET enzymatic activity to specific loci (***Liu et al., 2016***). 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 genes (***Sanz et al., 2016***; ***Chen et al., 2017***). 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.

5hmC 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 (***Chen et al., 2015***; ***Matarese et al., 2011***; ***Jin et al., 2014***; ***Nadel et al., 2015***). To assess individual genome-wide distribution profiles, R-loops density was probed over fixed windows of ±10 kbp around the 5hmC peaks (***Figure 3A***, ***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 ***Chen et al., 2015***), we could obtain a statistically significant Pearson correlation coefficient between both ($p<0.05$) (***Figure 3B***, ***Figure 3—figure supplement 1B***). Furthermore, approximately half of all R-loops detected genome-wide occurred at 5hmC-containing loci (***Figure 3C***, ***Figure 3—figure supplement 1C***). Notably, we observed an overlap between 5hmC and R-loops in 6839 (51%) out of the 13,288 actively expressed genes (***Figure 3D***), a feature illustrated in the individual profiles of mouse and

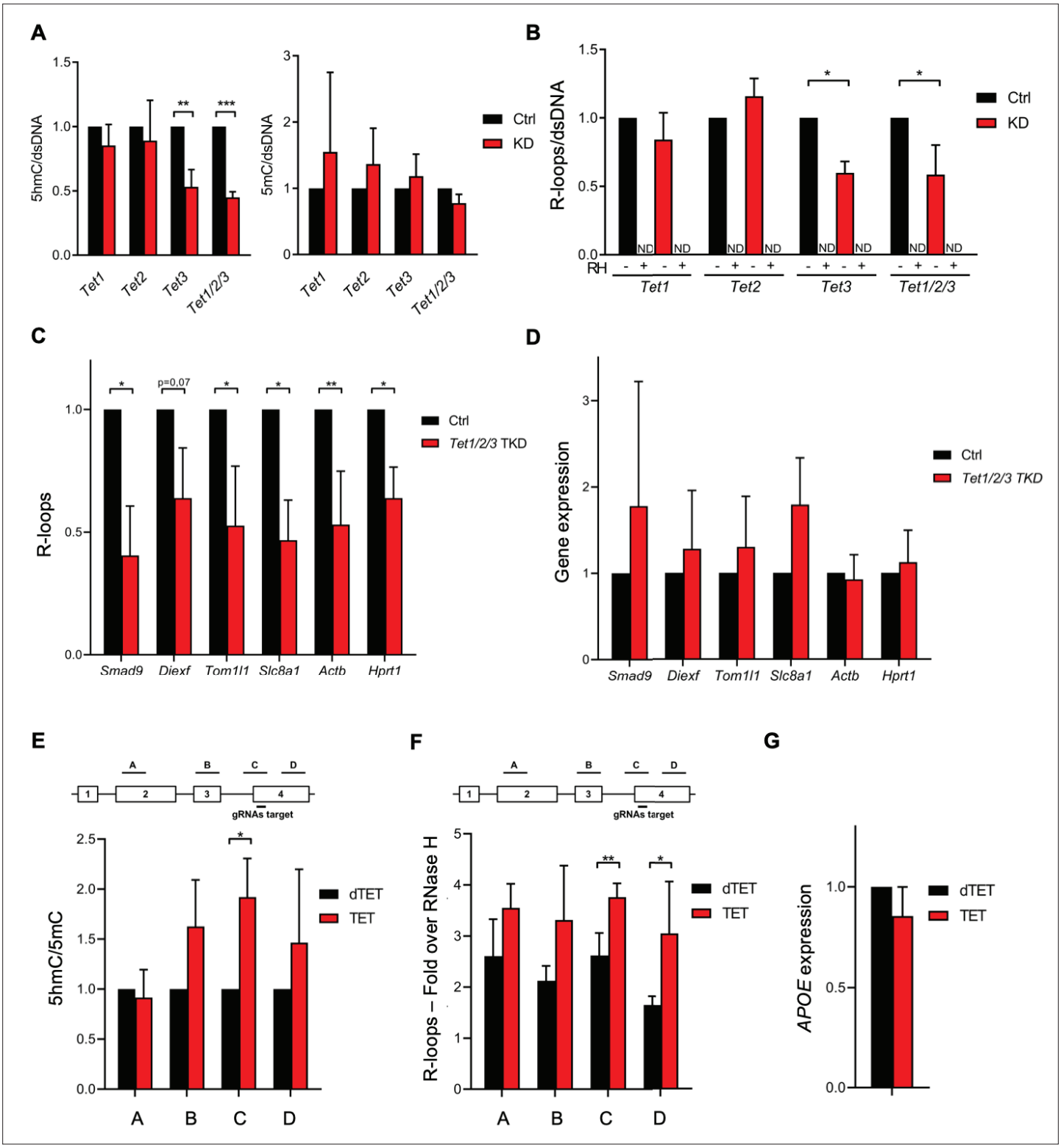


Figure 2. Ten-eleven translocation (TET) enzymatic activity impacts R-loop levels. Quantification of 5-hydroxymethylcytosine (5hmC) and 5-methylcytosine (5mC) (**A**) and R-loops (**B**) dot blots of *Tet1*, *Tet2*, *Tet3* single KD, and of *Tet1/2/3* triple KD mouse embryonic stem (mES) cells. Dot blots 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 DNA:RNA immunoprecipitation (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 four regions of the *APOE* gene upon tethering of dCas9-TET1 or dCas9-

Figure 2 continued on next page

Figure 2 continued

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.

The online version of this article includes the following figure supplement(s) for figure 2:

Figure supplement 1. Impact of *Tet* depletion in mouse embryonic stem (mES) cells in R-loops, 5-hydroxymethylcytosine (5hmC), and 5-methylcytosine (5mC).

Figure supplement 2. *Tet* depletion impacts R-loop formation in mouse fibroblasts.

human genes (***Figure 3E***, ***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 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 (***Ginno et al., 2013***; ***Ginno et al., 2012***) and implies 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 (***Nojima and Proudfoot, 2022***). 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 (PLAs) 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 (***García-Muse and Aguilera, 2019***). For instance, co-transcriptional R-loops increase conflicts between transcription and replication machineries by creating an additional barrier to fork progression (***Hamperl et al., 2017***; ***Helmrich et al., 2013***). 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 (***Hatchi et al., 2015***). We then sought to investigate if 5hmC creates conditions for DNA damage by promoting R-loop formation. We analyzed the genomic distribution of γH2AX by interrogating chromatin immunoprecipitation followed by sequencing (ChIP-seq) data from HEK293 cells (***Bunch et al., 2015***). The individual distribution profiles of γH2AX were analyzed 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 analyzed whole-transcriptome (RNA-seq) of mES cells overexpressing RNase H, a condition resulting in genome-wide loss of R-loops (***Chen et al., 2015***). 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) signaling pathways (***Figure 6B and C***). mTOR and MYC are known to play

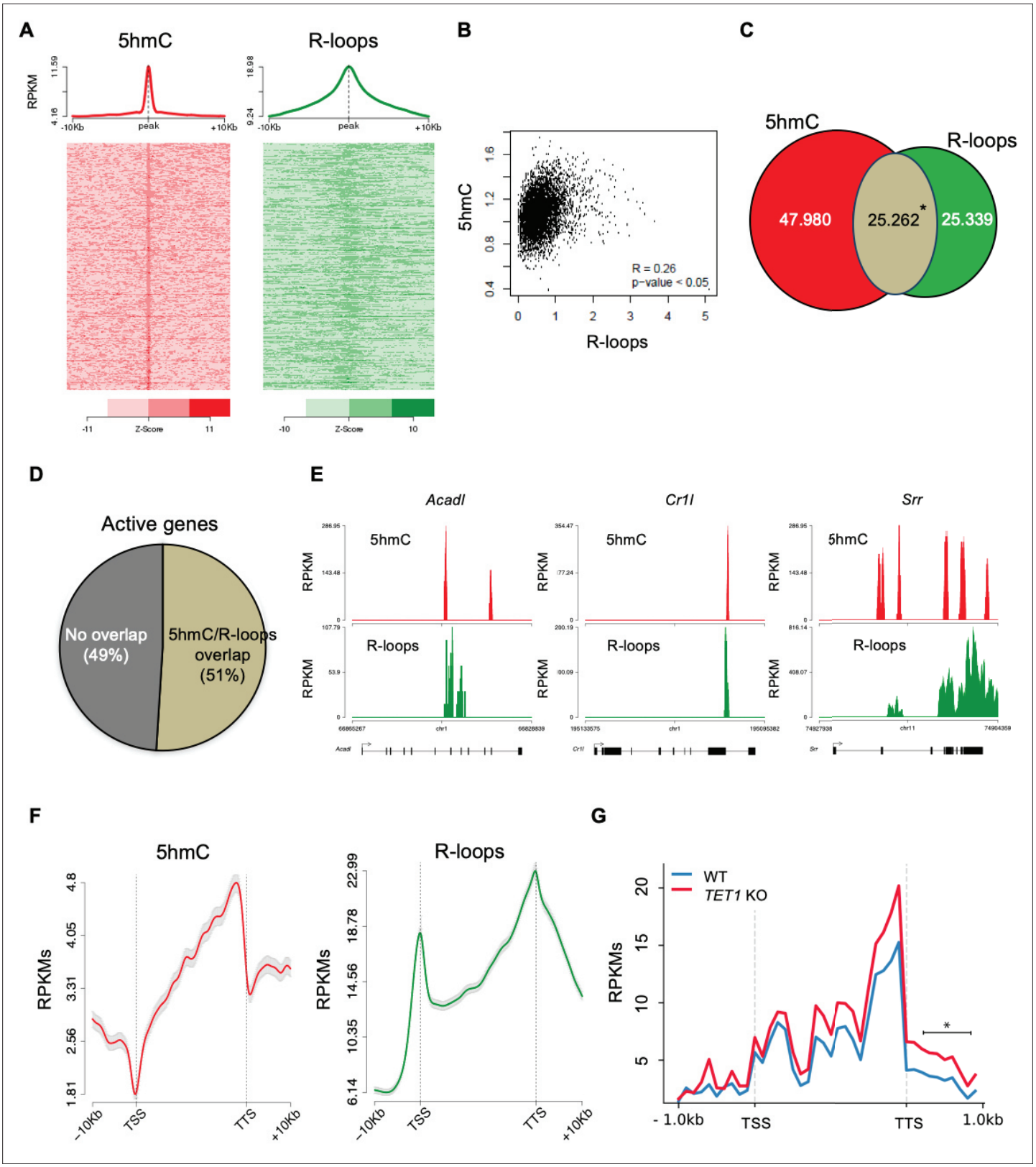


Figure 3. 5-Hydroxymethylcytosine (5hmC) and R-loops overlap in active genes of mouse embryonic stem (mES) cells. (**A**) Metagene and heatmap profiles of 5hmC and R-loops probed over fixed windows of ±10 kbp around the 5hmC peaks in expressed genes. (**B**) Pearson correlation coefficient between 5hmC 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

Figure 3 continued on next page

Figure 3 continued

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 200 bp 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.

The online version of this article includes the following figure supplement(s) for figure 3:

Figure supplement 1. Genome-wide analysis of 5-hydroxymethylcytosine (5hmC) and R-loops in HEK293 cells.

Figure supplement 2. Chromosomal distribution of R-loops and 5-hydroxymethylcytosine (5hmC).

opposite roles in establishing diapause, the temporary suspension of embryonic development driven by adverse environmental conditions (***Fenelon et al., 2014***), 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 (***Bulut-Karslioglu et al., 2016***). While overexpression of RNase H in mES cells did not reveal any significant changes in the cell cycle progression (***Figure 6—figure supplement 1A and 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 noncanonical DNA secondary structures, known as R-loops. Data from in vitro transcription reactions and AFM provide direct evidence showing that transcription through 5hmC-rich DNA favors R-loop formation. By depleting TET enzymes in mES cells and 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; ***García-Muse and Aguilera, 2019***) 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 suggests a putative causal link. Mechanistically, 5hmC may impact R-loop formation by either destabilizing the DNA duplex or altering RNA polymerase II elongation rate. Indeed, 5hmC modifies the DNA helix structure by favoring DNA-end breathing motion, diminishes the thermodynamic stability of the DNA duplex, and destabilizes GC pairing (***Mendonca et al., 2014***; ***Wanunu et al., 2011***). It also weakens the interaction between DNA and nucleosomal histones (***Mendonca et al., 2014***), which is thought to accelerate RNA polymerase II elongation but can also facilitate nascent RNA annealing with the template DNA strand favoring 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 (***García-Muse and Aguilera, 2019***), 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 (***García-Muse and Aguilera, 2019***), such as the regulation of gene expression. Our findings that genome-wide 5hmC and R-loops overlap robustly at the TTS 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 (***Skourti-Stathaki et al., 2011***).

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 (***Bray et al., 2021***; ***Good et al., 2018***). In addition to altering the

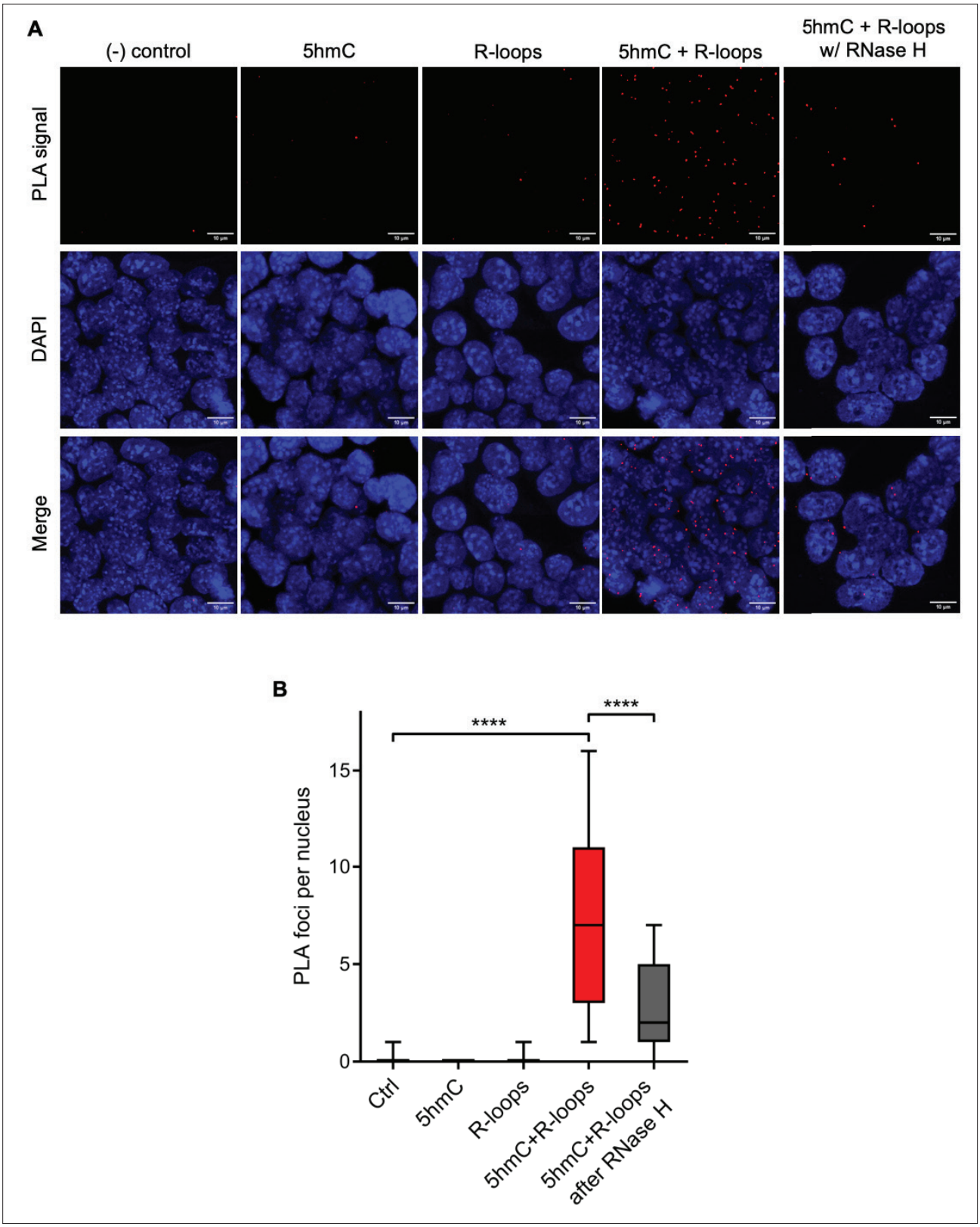


Figure 4. Simultaneous detection of 5-hydroxymethylcytosine (5hmC) and R-loops at the same genomic loci in individual mouse embryonic stem (mES) cells. (**A**) 5hmC and R-loops proximity ligation assay (PLA) foci in mES cells. DAPI was added to the mounting medium to stain DNA. Scale bars: 10 µm. Data 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.

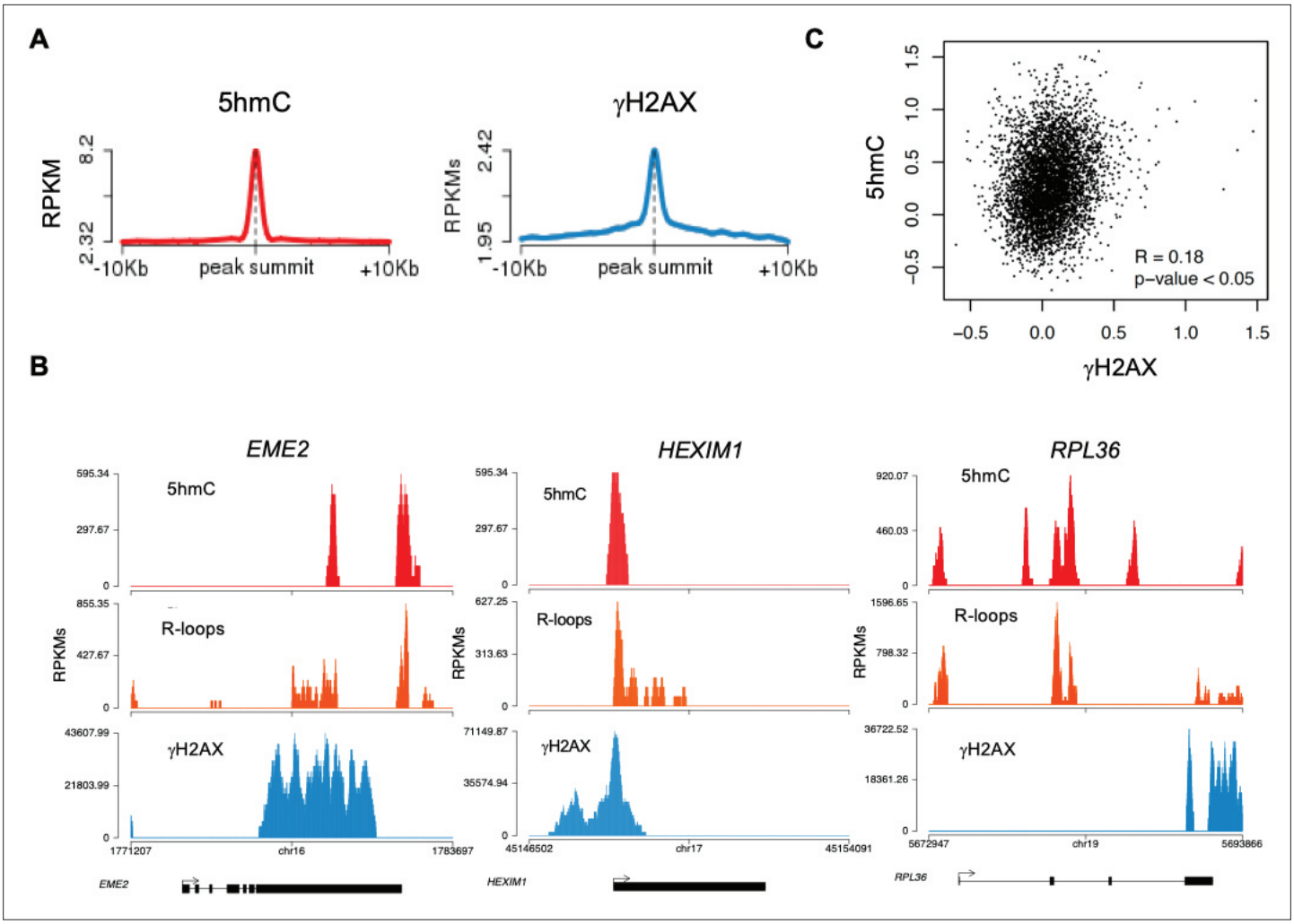


Figure 5. 5-hydroxymethylcytosine (5hmC)-rich loci are genomic hotspots for DNA damage. (**A**) Metagene profiles of 5hmC and γH2AX probed over fixed windows of ±10 kbp around the 5hmC peaks in expressed genes of HEK293 cells. (**B**) Individual profiles of 5hmC, R-loops and γH2AX distribution along 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 ($p<0.05$).

expression levels of tumour suppressors or oncogenes (***Bray et al., 2021***), 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 (***Good et al., 2017***), suggesting that mistargeted 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 (***Kafer et al., 2016***). 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 (***Ficz et al., 2011***). 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 (***Fenelon et al., 2014***). Accordingly, changes in the activity of mTOR, a major nutrient sensor, control ES cell commitment to trigger diapause (***Bulut-Karslioglu et al., 2016***). The mTOR signaling pathway was significantly downregulated upon global R-loop suppression by RNase H. Conversely, MYC targets,

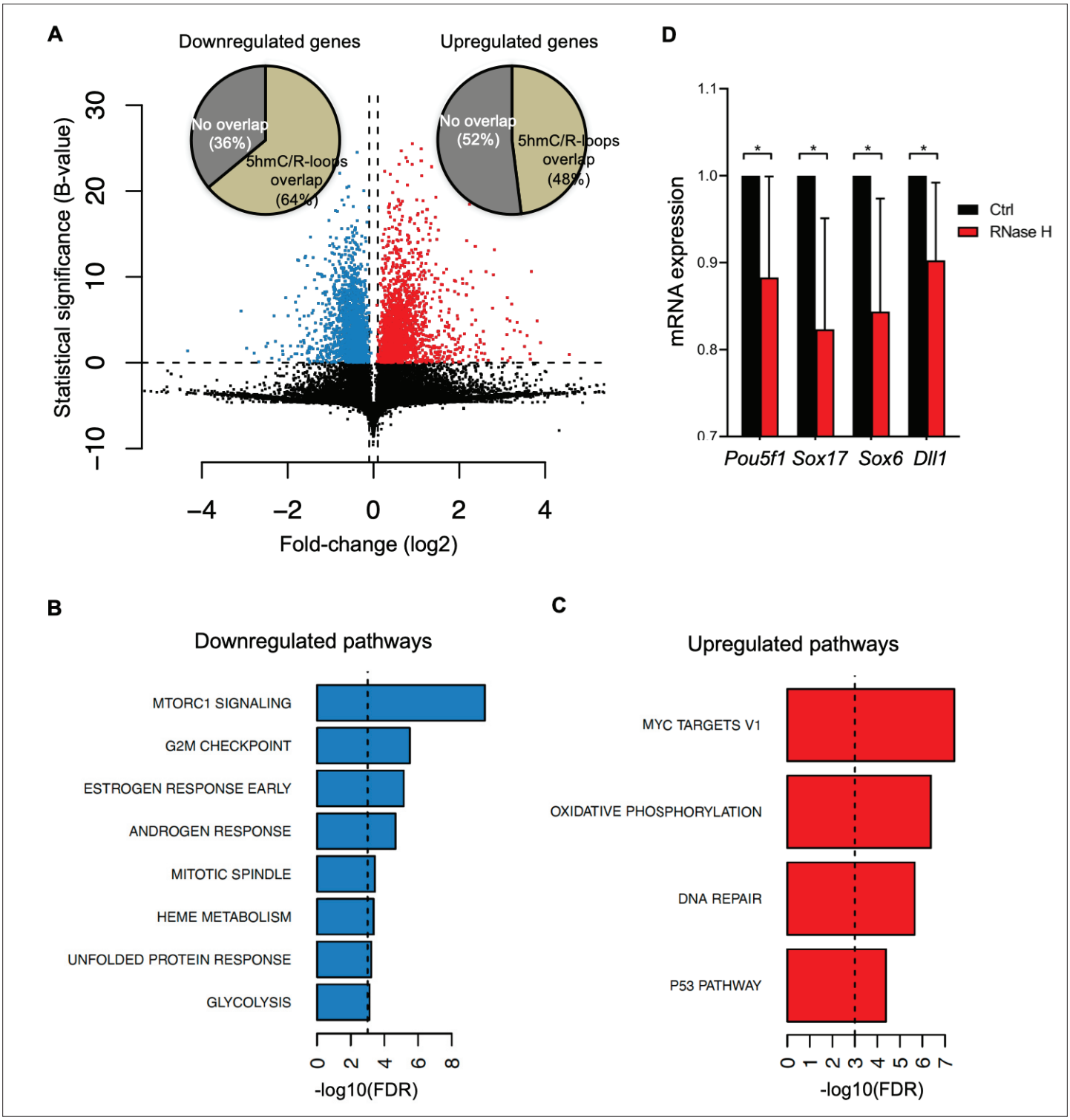


Figure 6. Cellular pathways affected by R-loops formed at 5-hydroxymethylcytosine (5hmC) loci. (**A**) Volcano plot displaying the differentially expressed genes in mouse embryonic stem (mES) cells upon RNase H overexpression. Of all downregulated and upregulated genes, 64 and 48% displayed R-loops overlapping 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.

The online version of this article includes the following figure supplement(s) for figure 6:

Figure supplement 1. Global R-loop suppression does not impact cell cycle progression of mouse embryonic stem (mES) cells.

which prevent ES cells from entering the state of dormancy that characterizes diapause (***Scognamiglio et al., 2016***), 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) mES cells were provided by Domingos Henrique (Instituto de Medicina Molecular João Lobo Antunes) and were a gift from Austin Smith (University of Exeter, UK) (***Smith and Hooper, 1987***). 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) 200 mM L-glutamine (25030-024, Thermo Scientific), 1% (v/v) 100 mM sodium pyruvate (11360-039, Gibco), 1% (v/v) 100× nonessential amino acids solution (11140-035, Gibco), 0.1% (v/v) 0.1 M 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 0,22 µm filter. Home-produced leukemia 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's medium (DMEM) (21969-035, Gibco), supplemented with 1% (v/v) 200 mM 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% CO_2. 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 four siRNAs provided as a single reagent was transfected using Lipofectamine RNAiMAX Transfection Reagent (13778150, Invitrogen) for 48 hr. 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 48 hr 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: $2^{(\text{Ct reference - Ct sample})}$, 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 (100 mM NaCl, 10 mM Tris pH 8.0, 25 mM 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 resuspended in DNase/RNase-free water. Nucleic acids were then fragmented using a restriction enzyme cocktail (20 U each of EcoRI, BamHI, HindIII, BsrgI, and XhoI). Half of the sample was digested with 40 U RNase H (MB085, NZYTech) for 48 hr 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 resuspended 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 1 hr at room temperature, and immunoblotted with specific antibodies. For the loading control, membranes were stripped in 0.5% SDS for 1 hr at 60°C, followed by blocking and re-probing. Details of the 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 10 min on ice, followed by 1 min acetone on ice. Cells were then incubated with primary antibodies for 1 hr 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 1 hr 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 3 U/µL RNase H (MB085, NZYTech) for 1 hr at 37°C prior to incubation with the antibodies. Images were acquired using the Point Scanning Confocal Microscope Zeiss LSM 880, 63×/1.4 oil immersion, with stacking acquisition and generation of maximum intensity projection images. PLA foci per nucleus were quantified using ImageJ. Details of the 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 the 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 the 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 2 hr at 37°C, using 1 µg of DNA as template, according to the 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 10 U RNase H (MB085, NZYTech) at 37°C overnight to serve as negative control. Then, all samples were treated with 0.05 U RNase A (10109142001, Roche) at 350 mM salt concentration for 15 min at 37°C 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 PBSt for 1 hr at room temperature, and incubated with the primary antibody at 4°C overnight. Signal quantification was performed using ImageJ. Details of the 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 resuspended in DNase/RNase-free water. DNA solution was diluted 1:10 in Sigma ultrapure water (with final 10 mM $MgCl_2$) and briefly mixed to ensure even dispersal in solution. A 10 µL droplet was deposited at the center 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 1–2 min 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. Afterward, the mica substrate was dried under a gentle stream of argon gas for approximately 2 min, 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 analyzed with JPK data processing software.

CRISPR-assisted 5mC/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 48 hr, cell culture supernatant was collected and filtered through a 0.45 µm filter. Lentivirus were collected through ultracentrifugation (25,000 rpm, 3 hr, 4°C) using an SW-41Ti rotor in a Beckman XL-90 ultracentrifuge. Virus were resuspended 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 24 hr, antibiotic selection was performed with 1.5 µg/mL puromycin, and infection proceeded for more 48 hr. 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 100 mM NaCl, 10 mM Tris pH 8.0, 25 mM EDTA, 0.5% SDS, 50 µg/mL Proteinase K overnight at 37°C. Nucleic acids were extracted using standard phenol-chloroform extraction protocol and resuspended in DNase/RNase-free water. Nucleic acids were then fragmented using a restriction enzyme cocktail (20 U each of EcoRI, BamHI, HindIII, BsrgI, and XhoI), and 10% of the digested sample was kept aside to use later as input. Half of the remaining volume was digested with 40 U RNase H (MB085, NZYTech) to serve as negative control, for 72 hr at 37°C. Digested nucleic acids were cleaned with standard phenol-chloroform extraction and resuspended in DNase/RNase-free water. DNA:RNA hybrids were immunoprecipitated from total nucleic acids using 5 µg of S9.6 antibody (MABE1095, Merck Millipore) in binding buffer (10 mM Na_2HPO_4 pH 7.0, 140 mM NaCl, 0.05% Triton X-100), overnight at 4°C. 50 µL protein G magnetic beads (10004D, Invitrogen) were used to pull down the immune complexes at 4°C for 2–3 hr. Isolated complexes were washed five times (for 1 min on ice) with binding buffer and once with Tris-EDTA (TE) buffer (10 mM Tris pH 8.1, 1 mM EDTA). Elution was performed in two steps, for 15 min at 55°C each, using elution buffer (50 mM Tris pH 8.0, 10 mM EDTA, 0.5% SDS, 60 µg/mL Proteinase K). The relative occupancy of DNA:RNA hybrids was estimated by RT-qPCR as follows: $2^{(Ct\ Input - Ct\ IP)}$, 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 100 mM NaCl, 10 mM Tris pH 8.0, 25 mM EDTA, 0.5% SDS, 50 µg/mL Proteinase K overnight at 37°C. Samples were sonicated with four pulses of 15 s at 10 mA intensity using a Soniprep150 sonicator (keeping tubes for at least 1 min on ice between pulses). Fragmented nucleic acids were cleaned with standard phenol-chloroform extraction protocol and resuspended 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 10 min, followed by immediate chilling on ice and quick spin. Samples were divided in half, and 5 µg of anti-5mC antibody (61255, Active Motif) or 5 µg of anti-5hmC antibody (39791, Active Motif) were used to immunoprecipitate 5mC and 5hmC, respectively, in binding buffer (10 mM Na_2HPO_4 pH 7.0, 140 mM NaCl, 0.05% Triton X-100), overnight at 4°C. 50 µL protein G magnetic beads (10004D, Invitrogen) were used to pull down the immune

complexes at 4°C for 2–3 hr. Isolated complexes were washed five times (for 1 min on ice) with binding buffer and once with TE buffer (10 mM Tris pH 8.1, 1 mM EDTA). Elution was performed in two steps, for 15 min at 55°C each, using elution buffer (50 mM Tris pH 8.0, 10 mM EDTA, 0.5% SDS, 60 µg/mL Proteinase K). The relative occupancy of 5mC and 5hmC was estimated by RT-qPCR as follows: $2^{(Ct\ Input - Ct\ IP)}$, 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 H1 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 48 hr later, cells were trypsinized and pelleted by centrifugation at 500 × *g* for 5 min. Cells were fixed in cold 1% PFA for 20 min at 4°C, followed by permeabilization in 70% ethanol for 1 hr at 4°C. Cells were then treated with 25 µg/mL RNase A (10109142001, Roche) in PBS 1× at 37°C for 20 min, 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 analyzed 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 (100 mM KAc, 30 mM 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 (https://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 analyzed through the same workflow. First, the reads were aligned to the reference mouse and human genome (mm10 and GRCh38/hg38 assemblies, respectively) with Bowtie (***Langmead et al., 2009***), and filtering for uniquely aligned reads. Enriched regions were identified relative to the input samples using MACS (***Zhang et al., 2008***), with a false discovery rate of 0.05. Finally, enriched regions were assigned to annotated genes, including a 4-kilobase region upstream the TSS and downstream the TTS. 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 (***Quinlan and Hall, 2010***). For individual and metaprofiles, uniquely mapped reads were extended in the 3′ direction to reach 150 nt with the Pyicos (***Althammer et al., 2011***). Individual profiles were produced using a 20 bp 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 200 bp 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 200 bp 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 (https://www.r-project.org/; ***R Core Team, 2018***). SAMtools (***Li et al., 2009***) 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 (***Bray et al., 2016***), 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 25th percentile. Differential expression in mES cells overexpressing RNase H was assessed using edgeR (v3.20.9) and limma (v3.34.9) R packages (***Robinson et al., 2010***; ***Ritchie et al., 2015***). Briefly, sample comparison was performed using voom transformed values, linear modeling, 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 downregulated 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) (***Liberzon et al., 2015***) 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 *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 (***Dobin et al., 2013***). Transcription readthrough levels were evaluated by counting the number of reads mapping downstream the TTS using ARTDeco (***Roth et al., 2020***) and human genome annotation from GENCODE project (GENCODE release 37). Genes with an enrichment in transcriptional readthrough in *TET1* KO samples relative to the control were identified. Metagene profiles were built using the *computeMatrix* tool from the deepTools v3.5.1 (***Ramírez et al., 2016***) and default packages from Python language. Genes were scaled to equally sized bins of 100 bp so that all annotated TSSs and TTSs were aligned. Regions of 1 kb were added upstream of TSS and downstream of TTS and also averaged in 100 bp bins. All read counts were normalized by the number of mapped reads (RPKM).

Acknowledgements

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). JCS is the recipient of an FCT PhD fellowship PD/BD/128292/2017. Work in CMA's laboratory is supported by "la Caixa" Foundation and FCT, IP (LCF/PR/HP21/52310016; PTDC/BIA-MOL/6624/2020; PTDC/MED-ONC/7864/2020).

Additional information

Funding

Funder	Grant reference number	Author
Fundação para a Ciência e Tecnologia, Portugal	PTDC/BIA-MOL/30438/2017	Sérgio Fernandes de Almeida
Fundação para a Ciência e Tecnologia, Portugal	PTDC/MED-OUT/4301/2020	Sérgio Fernandes de Almeida
EU Horizon 2020 Research and Innovation Programme	RiboMed 857119	Sérgio Fernandes de Almeida

Funder	Grant reference number	Author
Fundação para a Ciência e Tecnologia, Portugal	PD/BD/128292/2017	João C Sabino
La Caixa Foundation	LCF/PR/HP21/52310016	Claus M Azzalin
FCT	PTDC/BIA-MOL/6624/2020	Claus M Azzalin
FCT	PTDC/MED-ONC/7864/2020	Claus M Azzalin

The funders had no role in study design, data collection and interpretation, or the decision to submit the work for publication.

Author contributions

João C Sabino, Data curation, Formal analysis, Investigation, Methodology, Writing - review and editing; Madalena R de Almeida, Patrícia L Abreu, Marco M Domingues, Nuno C Santos, Claus M Azzalin, Formal analysis, Methodology, Writing - review and editing; Ana M Ferreira, Paulo Caldas, Formal analysis; Ana Rita Grosso, Formal analysis, Investigation, Methodology, Software, Writing - review and editing; Sérgio Fernandes de Almeida, Conceptualization, Formal analysis, Funding acquisition, Project administration, Supervision, Writing - original draft

Author ORCIDs

João C Sabino http://orcid.org/0000-0001-7991-4291
Madalena R de Almeida http://orcid.org/0000-0001-9539-3289
Patrícia L Abreu http://orcid.org/0000-0002-6387-8537
Claus M Azzalin http://orcid.org/0000-0002-9396-1980
Ana Rita Grosso http://orcid.org/0000-0001-6974-4209
Sérgio Fernandes de Almeida http://orcid.org/0000-0002-7774-1355

Decision letter and Author response

Decision letter https://doi.org/10.7554/eLife.69476.sa1
Author response https://doi.org/10.7554/eLife.69476.sa2

Additional files

Supplementary files

- Supplementary file 1. Differentially expressed genes upon RNase H overexpression.
- Supplementary file 2. Oligonucleotides and antibodies used in the study. (A) shRNA sequences. (B) Oligonucleotide sequences. (C) Antibodies used in this study. (D) g-Blocks sequences. (E) S9.6 electrophoretic mobility shift assay (EMSA) oligonucleotides.
- Transparent reporting form
- Source data 1. Original, uncropped images of blots and gels.
- Source data 2. Original, uncropped images of blots.

Data availability

All data generated or analysed during this study are included in the manuscript and supporting files.

The following previously published datasets were used:

Author(s)	Year	Dataset title	Dataset URL	Database and Identifier
Chen PB, Chen HV, Acharya D, Rando OJ	2015	R loops regulate promoter-proximal chromatin architecture and cellular differentiation	https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE67584	NCBI Gene Expression Omnibus, GSE67583

Continued on next page

Continued

Author(s)	Year	Dataset title	Dataset URL	Database and Identifier
Chen PB, Chen HV, Acharya D, Rando OJ	2015	Promoter-proximal R-loops regulate binding of chromatin regulators and pluripotency [DRIP-RNAseq]	https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE67581	NCBI Gene Expression Omnibus, GSE67581
Matarese F, Carrillo-de Santa Pau E, Stunnenberg HG	2011	5-hydroxymethylcytosine: the sixth DNA base	https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE31343	NCBI Gene Expression Omnibus, GSE31343
Flynn RA, Rubin AJ, Calo E, Bt DO	2016	7SK-BAF axis controls pervasive transcription at enhancers	https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE69140	NCBI Gene Expression Omnibus, GSE69140
Liu W, Ma Q, Wong K Li W	2014	Brd4 and JMJD6-associated anti-pause enhancers in regulation of transcriptional pause release	https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE51633	NCBI Gene Expression Omnibus, GSE51633
Nadel J, Athanasiadou R, Lemetre C, Wijetunga NA, ÓBroin P, Sato H, Zhang Z, Jeddeloh J, Montagna C, Golden A, Seoighe C, Greally J	2015	RNA:DNA hybrids in the human genome have distinctive nucleotide characteristics, chromatin composition, and transcriptional relationships	https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi	NCBI Gene Expression Omnibus, GSE68948
Jin C, Lu Y, Jelinek J, Liang S	2014	TET1 is a maintenance DNA demethylase that prevents methylation spreading in differentiated cells	https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE44036	NCBI Gene Expression Omnibus, GSE44036
Bunch H, Lawney BP, Lin YF, Asaithamby A	2015	Transcriptional elongation requires DNA break-induced signalling	https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE75170	NCBI Gene Expression Omnibus, GSE75170

References

Althammer S, González-Vallinas J, Ballaré C, Beato M, Eyras E. 2011. Pyicos: a versatile toolkit for the analysis of high-throughput sequencing data. *Bioinformatics (Oxford, England)* **27**:3333–3340. DOI: https://doi.org/10.1093/bioinformatics/btr570, PMID: 21994224

Arab K, Karaulanov E, Musheev M, Trnka P, Schäfer A, Grummt I, Niehrs C. 2019. GADD45A binds R-loops and recruits TET1 to CpG island promoters. *Nature Genetics* **51**:217–223. DOI: https://doi.org/10.1038/s41588-018-0306-6, PMID: 30617255

Bonnet A, Grosso AR, Elkaoutari A, Coleno E, Presle A, Sridhara SC, Janbon G, Géli V, de Almeida SF, Palancade B. 2017. Introns Protect Eukaryotic Genomes from Transcription-Associated Genetic Instability. *Molecular Cell* **67**:608–621. DOI: https://doi.org/10.1016/j.molcel.2017.07.002, PMID: 28757210

Bray NL, Pimentel H, Melsted P, Pachter L. 2016. Near-optimal probabilistic RNA-seq quantification. *Nature Biotechnology* **34**:525–527. DOI: https://doi.org/10.1038/nbt.3519, PMID: 27043002

Bray JK, Dawlaty MM, Verma A, Maitra A. 2021. Roles and Regulations of TET Enzymes in Solid Tumors. *Trends in Cancer* **7**:635–646. DOI: https://doi.org/10.1016/j.trecan.2020.12.011, PMID: 33468438

Bulut-Karslioglu A, Biechele S, Jin H, Macrae TA, Hejna M, Gertsenstein M, Song JS, Ramalho-Santos M. 2016. Inhibition of mTOR induces a paused pluripotent state. *Nature* **540**:119–123. DOI: https://doi.org/10.1038/nature20578, PMID: 27880763

Bunch H, Lawney BP, Lin Y-F, Asaithamby A, Murshid A, Wang YE, Chen BPC, Calderwood SK. 2015. Transcriptional elongation requires DNA break-induced signalling. *Nature Communications* **6**:10191. DOI: https://doi.org/10.1038/ncomms10191, PMID: 26671524

Carrasco-Salas Y, Malapert A, Sulthana S, Molcrette B, Chazot-Franguiadakis L, Bernard P, Chédin F, Faivre-Moskalenko C, Vanoosthuyse V. 2019. The extruded non-template strand determines the architecture of R-loops. *Nucleic Acids Research* **47**:6783–6795. DOI: https://doi.org/10.1093/nar/gkz341, PMID: 31066439

Chédin F. 2016. Nascent Connections: R-Loops and Chromatin Patterning. *Trends in Genetics* **32**:828–838. DOI: https://doi.org/10.1016/j.tig.2016.10.002, PMID: 27793359

Chen PB, Chen HV, Acharya D, Rando OJ, Fazzio TG. 2015. R loops regulate promoter-proximal chromatin architecture and cellular differentiation. *Nature Structural & Molecular Biology* **22**:999–1007. DOI: https://doi.org/10.1038/nsmb.3122, PMID: 26551076

Chen L, Chen JY, Zhang X, Gu Y, Xiao R, Shao C, Tang P, Qian H, Luo D, Li H, Zhou Y, Zhang DE, Fu XD. 2017. R-ChIP Using Inactive RNase H Reveals Dynamic Coupling of R-loops with Transcriptional Pausing at Gene Promoters. *Molecular Cell* **68**:745–757. DOI: https://doi.org/10.1016/j.molcel.2017.10.008, PMID: 29104020

Dobin A, Davis CA, Schlesinger F, Drenkow J, Zaleski C, Jha S, Batut P, Chaisson M, Gingeras TR. 2013. STAR: ultrafast universal RNA-seq aligner. *Bioinformatics (Oxford, England)* **29**:15–21. DOI: https://doi.org/10.1093/bioinformatics/bts635, PMID: 23104886

Fenelon JC, Banerjee A, Murphy BD. 2014. Embryonic diapause: development on hold. *The International Journal of Developmental Biology* **58**:163–174. DOI: https://doi.org/10.1387/ijdb.140074bm, PMID: 25023682

Ficz G, Branco MR, Seisenberger S, Santos F, Krueger F, Hore TA, Marques CJ, Andrews S, Reik W. 2011. Dynamic regulation of 5-hydroxymethylcytosine in mouse ES cells and during differentiation. *Nature* **473**:398–402. DOI: https://doi.org/10.1038/nature10008, PMID: 21460836

García-Muse T, Aguilera A. 2019. R Loops: From Physiological to Pathological Roles. *Cell* **179**:604–618. DOI: https://doi.org/10.1016/j.cell.2019.08.055, PMID: 31607512

Ginno PA, Lott PL, Christensen HC, Korf I, Chédin F. 2012. R-loop formation is a distinctive characteristic of unmethylated human CpG island promoters. *Molecular Cell* **45**:814–825. DOI: https://doi.org/10.1016/j.molcel.2012.01.017, PMID: 22387027

Ginno PA, Lim YW, Lott PL, Korf I, Chédin F. 2013. GC skew at the 5′ and 3′ ends of human genes links R-loop formation to epigenetic regulation and transcription termination. *Genome Research* **23**:1590–1600. DOI: https://doi.org/10.1101/gr.158436.113, PMID: 23868195

Good CR, Madzo J, Patel B, Maegawa S, Engel N, Jelinek J, Issa JPJ. 2017. A novel isoform of TET1 that lacks A CXXC domain is overexpressed in cancer. *Nucleic Acids Research* **45**:8269–8281. DOI: https://doi.org/10.1093/nar/gkx435, PMID: 28531272

Good CR, Panjarian S, Kelly AD, Madzo J, Patel B, Jelinek J, Issa JPJ. 2018. TET1-Mediated Hypomethylation Activates Oncogenic Signaling in Triple-Negative Breast Cancer. *Cancer Research* **78**:4126–4137. DOI: https://doi.org/10.1158/0008-5472.CAN-17-2082, PMID: 29891505

Greenberg MVC, Bourc'his D. 2019. The diverse roles of DNA methylation in mammalian development and disease. *Nature Reviews. Molecular Cell Biology* **20**:590–607. DOI: https://doi.org/10.1038/s41580-019-0159-6, PMID: 31399642

Grunseich C, Wang IX, Watts JA, Burdick JT, Guber RD, Zhu Z, Bruzel A, Lanman T, Chen K, Schindler AB, Edwards N, Ray-Chaudhury A, Yao J, Lehky T, Piszczek G, Crain B, Fischbeck KH, Cheung VG. 2018. Senataxin Mutation Reveals How R-Loops Promote Transcription by Blocking DNA Methylation at Gene Promoters. *Molecular Cell* **69**:426–437. DOI: https://doi.org/10.1016/j.molcel.2017.12.030, PMID: 29395064

Hahn MA, Qiu R, Wu X, Li AX, Zhang H, Wang J, Jui J, Jin S-G, Jiang Y, Pfeifer GP, Lu Q. 2013. Dynamics of 5-hydroxymethylcytosine and chromatin marks in Mammalian neurogenesis. *Cell Reports* **3**:291–300. DOI: https://doi.org/10.1016/j.celrep.2013.01.011, PMID: 23403289

Hamperl S, Bocek MJ, Saldivar JC, Swigut T, Cimprich KA. 2017. Transcription-Replication Conflict Orientation Modulates R-Loop Levels and Activates Distinct DNA Damage Responses. *Cell* **170**:774–786. DOI: https://doi.org/10.1016/j.cell.2017.07.043, PMID: 28802045

Hatchi E, Skourti-Stathaki K, Ventz S, Pinello L, Yen A, Kamieniarz-Gdula K, Dimitrov S, Pathania S, McKinney KM, Eaton ML, Kellis M, Hill SJ, Parmigiani G, Proudfoot NJ, Livingston DM. 2015. BRCA1 recruitment to transcriptional pause sites is required for R-loop-driven DNA damage repair. *Molecular Cell* **57**:636–647. DOI: https://doi.org/10.1016/j.molcel.2015.01.011, PMID: 25699710

Helmrich A, Ballarino M, Nudler E, Tora L. 2013. Transcription-replication encounters, consequences and genomic instability. *Nature Structural & Molecular Biology* **20**:412–418. DOI: https://doi.org/10.1038/nsmb.2543, PMID: 23552296

Jin C, Lu Y, Jelinek J, Liang S, Estecio MRH, Barton MC, Issa J-PJ. 2014. TET1 is a maintenance DNA demethylase that prevents methylation spreading in differentiated cells. *Nucleic Acids Research* **42**:6956–6971. DOI: https://doi.org/10.1093/nar/gku372, PMID: 24875481

Kafer GR, Li X, Horii T, Suetake I, Tajima S, Hatada I, Carlton PM. 2016. 5-Hydroxymethylcytosine Marks Sites of DNA Damage and Promotes Genome Stability. *Cell Reports* **14**:1283–1292. DOI: https://doi.org/10.1016/j.celrep.2016.01.035, PMID: 26854228

Karpf AR. 2013. Epigenetic alterations in oncogenesis. Preface. *Advances in Experimental Medicine and Biology* **754**:v–vii. DOI: https://doi.org/10.1007/978-1-4419-9967-2, PMID: 23189391

Klinov DV, Lagutina IV, Prokhorov VV, Neretina T, Khil PP, Lebedev YB, Cherny DI, Demin VV, Sverdlov ED. 1998. High resolution mapping DNAs by R-loop atomic force microscopy. *Nucleic Acids Research* **26**:4603–4610. DOI: https://doi.org/10.1093/nar/26.20.4603, PMID: 9753727

Langmead B, Trapnell C, Pop M, Salzberg SL. 2009. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biology* **10**:R25. DOI: https://doi.org/10.1186/gb-2009-10-3-r25, PMID: 19261174

Leavitt R, Yen J, Jia XY. 2015. 5-methylcytosine and 5-hydroxymethylcytosine Exert Opposite Forces on Base Pairing of DNA Double Helix. *Zymo Research Corporation* **1**:6–7.

Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin R, 1000 Genome Project Data Processing Subgroup. 2009. The Sequence Alignment/Map format and SAMtools. *Bioinformatics (Oxford, England)* **25**:2078–2079. DOI: https://doi.org/10.1093/bioinformatics/btp352, PMID: 19505943

Liberzon A, Birger C, Thorvaldsdóttir H, Ghandi M, Mesirov JP, Tamayo P. 2015. The Molecular Signatures Database (MSigDB) hallmark gene set collection. *Cell Systems* **1**:417–425. DOI: https://doi.org/10.1016/j.cels.2015.12.004, PMID: 26771021

Liu XS, Wu H, Ji X, Stelzer Y, Wu X, Czauderna S, Shu J, Dadon D, Young RA, Jaenisch R. 2016. Editing DNA Methylation in the Mammalian Genome. *Cell* **167**:233–247. DOI: https://doi.org/10.1016/j.cell.2016.08.056, PMID: 27662091

Matarese F, Carrillo-de Santa Pau E, Stunnenberg HG. 2011. 5-Hydroxymethylcytosine: a new kid on the epigenetic block? *Molecular Systems Biology* **7**:562. DOI: https://doi.org/10.1038/msb.2011.95, PMID: 22186736

Mendonca A, Chang EH, Liu W, Yuan C. 2014. Hydroxymethylation of DNA influences nucleosomal conformation and stability in vitro. *Biochimica et Biophysica Acta* **1839**:1323–1329. DOI: https://doi.org/10.1016/j.bbagrm.2014.09.014, PMID: 25263161

Nadel J, Athanasiadou R, Lemetre C, Wijetunga NA, Ó Broin P, Sato H, Zhang Z, Jeddeloh J, Montagna C, Golden A, Seoighe C, Greally JM. 2015. RNA:DNA hybrids in the human genome have distinctive nucleotide characteristics, chromatin composition, and transcriptional relationships. *Epigenetics & Chromatin* **8**:46. DOI: https://doi.org/10.1186/s13072-015-0040-6, PMID: 26579211

Nojima T, Proudfoot NJ. 2022. Mechanisms of lncRNA biogenesis as revealed by nascent transcriptomics. *Nature Reviews. Molecular Cell Biology* **1**:0123456789. DOI: https://doi.org/10.1038/s41580-021-00447-6, PMID: 35079163

Pastor WA, Aravind L, Rao A. 2013. TETonic shift: biological roles of TET proteins in DNA demethylation and transcription. *Nature Reviews. Molecular Cell Biology* **14**:341–356. DOI: https://doi.org/10.1038/nrm3589, PMID: 23698584

Quinlan AR, Hall IM. 2010. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics (Oxford, England)* **26**:841–842. DOI: https://doi.org/10.1093/bioinformatics/btq033, PMID: 20110278

R Core Team. 2018. R: A Language and Environment for Statistical Computing. Austria: R Foundation for Statistical Computing.

Ramírez F, Ryan DP, Grüning B, Bhardwaj V, Kilpert F, Richter AS, Heyne S, Dündar F, Manke T. 2016. deepTools2: a next generation web server for deep-sequencing data analysis. *Nucleic Acids Research* **44**:W160–W165. DOI: https://doi.org/10.1093/nar/gkw257

Ritchie ME, Phipson B, Wu D, Hu Y, Law CW, Shi W, Smyth GK. 2015. limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic Acids Research* **43**:e47. DOI: https://doi.org/10.1093/nar/gkv007, PMID: 25605792

Robinson MD, McCarthy DJ, Smyth GK. 2010. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics (Oxford, England)* **26**:139–140. DOI: https://doi.org/10.1093/bioinformatics/btp616, PMID: 19910308

Roth SJ, Heinz S, Benner C. 2020. ARTDeco: automatic readthrough transcription detection. *BMC Bioinformatics* **21**:1–22. DOI: https://doi.org/10.1186/s12859-020-03551-0

Sanz LA, Hartono SR, Lim YW, Steyaert S, Rajpurkar A, Ginno PA, Xu X, Chédin F. 2016. Prevalent, Dynamic, and Conserved R-Loop Structures Associate with Specific Epigenomic Signatures in Mammals. *Molecular Cell* **63**:167–178. DOI: https://doi.org/10.1016/j.molcel.2016.05.032, PMID: 27373332

Scognamiglio R, Cabezas-Wallscheid N, Thier MC, Altamura S, Reyes A, Prendergast ÁM, Baumgärtner D, Carnevalli LS, Atzberger A, Haas S, von Paleske L, Boroviak T, Wörsdörfer P, Essers MAG, Kloz U, Eisenman RN, Edenhofer F, Bertone P, Huber W, van der Hoeven F, et al. 2016. Myc Depletion Induces a Pluripotent Dormant State Mimicking Diapause. *Cell* **164**:668–680. DOI: https://doi.org/10.1016/j.cell.2015.12.033, PMID: 26871632

Skourti-Stathaki K, Proudfoot NJ, Gromak N. 2011. Human senataxin resolves RNA/DNA hybrids formed at transcriptional pause sites to promote Xrn2-dependent termination. *Molecular Cell* **42**:794–805. DOI: https://doi.org/10.1016/j.molcel.2011.04.026, PMID: 21700224

Skourti-Stathaki K, Kamieniarz-Gdula K, Proudfoot NJ. 2014. R-loops induce repressive chromatin marks over mammalian gene terminators. *Nature* **516**:436–439. DOI: https://doi.org/10.1038/nature13787, PMID: 25296254

Skourti-Stathaki K, Proudfoot NJ. 2014. A double-edged sword: R loops as threats to genome integrity and powerful regulators of gene expression. *Genes & Development* **28**:1384–1396. DOI: https://doi.org/10.1101/gad.242990.114, PMID: 24990962

Smith AG, Hooper ML. 1987. Buffalo rat liver cells produce a diffusible activity which inhibits the differentiation of murine embryonal carcinoma and embryonic stem cells. *Developmental Biology* **121**:1–9. DOI: https://doi.org/10.1016/0012-1606(87)90132-1, PMID: 3569655

Sridhara SC, Carvalho S, Grosso AR, Gallego-Paez LM, Carmo-Fonseca M, de Almeida SF. 2017. Transcription Dynamics Prevent RNA-Mediated Genomic Instability through SRPK2-Dependent DDX23 Phosphorylation. *Cell Reports* **18**:334–343. DOI: https://doi.org/10.1016/j.celrep.2016.12.050, PMID: 28076779

Tahiliani M, Koh KP, Shen Y, Pastor WA, Bandukwala H, Brudno Y, Agarwal S, Iyer LM, Liu DR, Aravind L, Rao A. 2009. Conversion of 5-methylcytosine to 5-hydroxymethylcytosine in mammalian DNA by MLL partner TET1. *Science (New York, N.Y.)* **324**:930–935. DOI: https://doi.org/10.1126/science.1170116, PMID: 19372391

Wanunu M, Cohen-Karni D, Johnson RR, Fields L, Benner J, Peterman N, Zheng Y, Klein ML, Drndic M. 2011. Discrimination of methylcytosine from hydroxymethylcytosine in DNA molecules. *Journal of the American Chemical Society* **133**:486–492. DOI: https://doi.org/10.1021/ja107836t, PMID: 21155562

Weber M, Hellmann I, Stadler MB, Ramos L, Pääbo S, Rebhan M, Schübeler D. 2007. Distribution, silencing potential and evolutionary impact of promoter DNA methylation in the human genome. *Nature Genetics* **39**:457–466. DOI: https://doi.org/10.1038/ng1990, PMID: 17334365

Wu H, D'Alessio AC, Ito S, Wang Z, Cui K, Zhao K, Sun YE, Zhang Y. 2011. Genome-wide analysis of 5-hydroxymethylcytosine distribution reveals its dual function in transcriptional regulation in mouse embryonic stem cells. *Genes & Development* **25**:679–684. DOI: https://doi.org/10.1101/gad.2036011, PMID: 21460036

Zhang Y, Liu T, Meyer CA, Eeckhoute J, Johnson DS, Bernstein BE, Nusbaum C, Myers RM, Brown M, Li W, Liu XS. 2008. Model-based analysis of ChIP-Seq (MACS). *Genome Biology* **9**:R137. DOI: https://doi.org/10.1186/gb-2008-9-9-r137, PMID: 18798982