New methods to study DNA methylation and the insights from their application

by
Daniel Mark Sapozhnikov

Department of Pharmacology and Therapeutics
Faculty of Medicine and Health Sciences

McGill University, Montreal

July 2023

A thesis submitted to McGill University in partial fulfillment of the requirements of the degree of Doctor of Philosophy in Pharmacology and Therapeutics

Table of Contents

Thesis Format

This is a manuscript-based thesis which conforms to the “Guidelines for Thesis Preparation” specified by the Graduate and Postdoctoral Studies unit at McGill University. The thesis is composed of four chapters and the manuscripts that make up the body of the thesis are arranged in the sequence of their publication or submission for publication. Chapter 1 provides a comprehensive review of the literature as it relates to the subjects discussed in the body of the thesis. Sections 11.2.3-11.2.5 of Chapter 1 contain portions of text and figures that were published in *Biomedicines*[1]. Chapter 2 is a manuscript that was published in *Nature Communications*[2] and presents a method for targeted DNA demethylation with CRISPR/dCas9 and its use to explore the relationship between DNA demethylation and gene expression. Chapter 3 is a manuscript that is, at the time of writing, under review at *The EMBO Journal* and presents a method for the sequencing of epigenetic modifications and its use in interrogating the dynamics of the active DNA demethylation pathway. Finally, Chapter 4 represents a general discussion of the thesis and includes portions of text that were published in *Biomedicines*[1] and in *Nature Protocols*[3].

Abstract

DNA methylation and demethylation are essential processes that play a pivotal role in regulating gene expression and maintaining genomic stability. DNA methylation involves the enzymatic addition of a methyl group to cytosine bases of DNA, primarily in the context of CpG dinucleotides. When this modification occurs in promoters, it typically leads to gene silencing. On the other hand, active DNA demethylation refers to the enzymatic removal of methyl groups from DNA, allowing for re-activation of gene expression. Together, these dynamic processes orchestrate a complex regulation of gene expression and thereby influence diverse biological phenomena, including embryonic development, cellular differentiation, genomic imprinting, X-chromosome inactivation, behavior, and cancer progression. Understanding the mechanisms and functions of DNA methylation and active DNA demethylation is crucial for unraveling the intricacies of gene regulation and its impact on various physiological and pathological processes. The current understanding of the importance of specific instances of DNA methylation in physiological processes and diseases is largely based on correlational studies across which researchers have observed associations between DNA methylation patterns and gene expression, disease states, or environmental factors. While correlational data can provide valuable insights into potential roles of DNA methylation, they provide no information as to the causal relationship between the change in methylation and the change in expression, especially in the context of a complex and dynamic nuclear environment that produces changes in gene expression as the convergent output of numerous processes that include other epigenetic modifications, chromatin structure, and transcription factor binding. The active DNA demethylation pathway, on the other hand, was only discovered approximately a decade ago and, therefore, the consequences on gene expression exerted by its protein components and its reaction intermediates – beyond the presumed effects of only demethylation – have not yet been studied in detail.

In the work presented in this thesis, I developed two technologies in an effort to address these challenges and expand the understanding of DNA methylation and demethylation. Chapter 2 describes the development, optimization, and implementation of a CRISPR/Cas9-based targeted methylation editing technique that relies on the simple physical interference of a dCas9 protein with DNA methyltransferase activity at targeted sites and, thus, causes demethylation in dividing cells without any confounding epigenetic activity. I show that this approach achieves efficient and specific demethylation and can be used to study methylation in different contexts, thus representing a useful new method to modify DNA methylation levels at precise genomic locations in live cells and enable researchers to investigate the causal relationship between DNA demethylation and gene expression across diverse contexts. In Chapter 3, I report a simple and efficient technique for the sequencing of the oxidized cytosine intermediates of the active DNA demethylation pathway, enabling the ability to discriminate oxidized cytosines from unmethylated cytosines at single-base resolution. I use this technique in combination with numerous genetic perturbations to dissect the active DNA demethylation pathway. The results of these experiments provide novel insights into the functions and interactions of the proteins and oxidized cytosine bases involved in this pathway. In summary, this thesis presents two new tools to study both passive and active DNA demethylation and reports several findings from their initial application with the primary goal that these methods simplify and improve future research into DNA methylation and demethylation.

Résumé

La méthylation de l'ADN implique l'ajout enzymatique d'un groupe méthyle aux bases cytosine de l'ADN, principalement dans le contexte des dinucléotides CpG. Lorsque cette modification se produit au niveau des promoteurs, elle entraîne généralement l'inhibition de l'expression des gènes. D'autre part, la déméthylation active de l'ADN fait référence à l'élimination enzymatique des groupes méthyles de l'ADN, ce qui permet de réactiver l'expression des gènes. La compréhension des mécanismes et des fonctions de la méthylation de l'ADN et de la déméthylation active de l'ADN est cruciale pour élucider les subtilités de la régulation des gènes et son impact sur divers processus physiologiques et pathologiques. Notre compréhension actuelle de l'importance de certains cas de méthylation de l'ADN dans les processus physiologiques et les maladies repose en grande partie sur des études corrélationnelles dans lesquelles les chercheurs ont observé des associations entre les profils de méthylation de l'ADN et l'expression des gènes, les états pathologiques ou les facteurs environnementaux. Même si les données corrélationnelles peuvent fournir des indications précieuses sur les rôles potentiels de la méthylation de l'ADN, elles ne fournissent aucune information sur la relation de cause à effet entre le changement de méthylation et le changement d'expression. De plus, dans le contexte d'un environnement complexe et dynamique du noyau cellulaire produisant des changements dans l'expression des gènes, ces effets résultent de nombreux processus comprenant d'autres modifications épigénétiques, la structure de la chromatine et la liaison des facteurs de transcription. En revanche, la voie active de déméthylation de l'ADN n'a été découverte qu'il y a une dizaine d'années et, par conséquent, les conséquences sur l'expression des gènes exercées par ses composants protéiques et ses intermédiaires réactionnels n'ont pas encore été étudiées en détail, au-delà des effets présumés uniquement par la déméthylation.

À travers le travail présenté dans cette thèse, j'ai développé deux technologies ayant pour but de relever ces défis et d'approfondir la compréhension de la méthylation et de la déméthylation de l'ADN. Le chapitre 2 décrit le développement, l'optimisation et la

mise en œuvre d'une technique d'édition ciblée de la méthylation basée sur CRISPR/Cas9 qui repose sur la simple interférence physique entre une protéine dCas9 et l'activité de l'ADN méthyltransférase sur des sites ciblés. Par conséquent, cette approche provoque une déméthylation dans les cellules en division sans aucune activité épigénétique confondante. Je démontre que cette approche permet une déméthylation efficace et spécifique et peut être utilisée pour étudier la méthylation dans différents contextes. Cette technologie représente une nouvelle méthode utile pour modifier les niveaux de méthylation de l'ADN à des endroits précis du génome dans des cellules vivantes et permet aux chercheurs d'étudier la relation de cause à effet entre la déméthylation de l'ADN et l'expression des gènes dans divers contextes. Dans le chapitre 3, je présente une technique simple et efficace pour le séquençage des cytosines oxydées intermédiaires de la voie active de déméthylation de l'ADN, qui permet de distinguer les cytosines oxydées des cytosines non méthylées à une résolution d'une seule base. J'utilise cette technique en combinaison avec de nombreuses perturbations génétiques pour disséquer la voie active de déméthylation de l'ADN. Les résultats de ces expériences fournissent de nouvelles connaissances sur les fonctions et les interactions des protéines et des cytosines oxydées impliquées dans cette voie. En résumé, cette thèse présente deux nouveaux outils pour étudier la déméthylation passive et active de l'ADN et rapporte plusieurs résultats de leur application initiale.

Acknowledgements

I am so undoubtedly fortunate to have reached this final stage of my education and to have so many magnificent people to thank for their mentorship, their help, their support, their belief and trust in me, their love, their companionship, and their acts of kindness, large, small, constant, or even the most miniscule, effortless, and momentary. And yet, I must ask all of you to forgive me for not beginning here with the detailing of these very dear acknowledgements, because the only gratitude I can give in this terribly bittersweet moment is to my father, Michael Sapozhnikov…to most, Misha…to me, Papa…who passed away – so unexpectedly, so suddenly and without a goodbye, and so incredibly too soon – as I whittled away at the contents of this thesis. It is you, Papa, to whom I owe my greatest qualities.

My father taught himself to do every single thing he ever decided that he needed – whether electrical wiring in the walls or tiling the floor or new computer programming languages or underground drainage systems in our backyard and so, so much more – through hundreds of printouts and diagrams and notes scribbled on index cards and organized into neat little labeled folders and boxes. Across all these things he did, he insisted on doing them slowly, with the greatest care for every minute detail, so that everything that he produced was of the highest quality. And, through every task we did together, he soaked my own fibers in those principles – most intentionally – and, in doing so, drove me to find every small success I have achieved. Though he was certainly not a scientist, these are the same qualities that enabled me to learn countless scientific concepts and techniques during the years I spent as the only student in the laboratory and to strive to perform these techniques at the highest levels; qualities which, I think, would serve anyone in any role. And so, every success I share within this thesis and without was compounded by every most miniscule task done thoroughly, to the best of my ability, with the utmost care, not settling until it was done well, no more and no less than as he taught me.

He was so excited to come to my graduation, even when I repeatedly tried to disenchant him from the idea, telling him I might be too busy – working, perhaps, in a different city – and he had seen me wear the same clothes, stand in the same tent, on the same campus when I received my Bachelor's degree. And we would certainly have celebrated in the very same restaurant, too. Now, I have no choice but to try to find some happiness in the fact that he *did* experience a very similar scene and we *did* get to share it together, our photos that day showing him boast some of the biggest smiles I have ever seen him wear. It will forever be my deepest sorrow for him to have never seen these moments now, nor the way my life thereafter unfolds – not to be able to share these moments with him – and there is naught to say but that life is cruel and unfair and, while I had heard those words so many times, I know now that, in happy ignorance, I had not the slightest clue as to what they meant.

In this alternate reality that I find myself in him, I see no better use for this small section amidst the exhaustive details of seven years of my work than to publicly and permanently express my gratitude to my father for all he did for me and all he taught me. So, to my father – and to my mother, who did it all alongside him – this was all thanks to you and this – and every single thing that follows – is all for you. And with that in mind it is with the same sincerity and admiration that I also deeply thank my mother, Yelena Sinitsyn, who, in addition to the amazing example she has been, has always put her greatest efforts into caring for me, guiding me through all decisions, affording me every opportunity, and sculpting me to have the right motivations and principles to be decent, considerate, and hard-working. There are no words to convey the depth of my respect and gratitude for all that you have done and continue to do.

Of course, I must give thanks where thanks is due – where it has been so sincerely earned – and it is due in its largest piece to my PhD supervisor, Dr. Moshe Szyf. I came into his office as an undergraduate looking for my first taste of research experience and he not only granted me an absolutely thrilling capacity to play with and alter biology as if

it truly were just dough in my hands, but also to do so within the framework of an exciting project which I helped to conceive, with no one to report to or to guide me besides Moshe. So, in this role, he introduced me to and steered me through the smallest technical details, caveats, and flaws of a wide landscape of scientific methodology. He designed experiments with me, shared in my excitement, and when the many disappointments of experimental work took some of that excitement from me, he always ignited it again. He allowed me the highest levels of independence and placed the greatest trust in my abilities across all scientific endeavours. He taught me how, in the context of science and beyond, to be critical but to also open my mind. He set the greatest example of a deep and excellent scientific thinker with a true love for science. I will always be thankful for everything he did to foster the scientific success that I have had and, beyond that, all of the hundreds of thoughtful and philosophical conversations we have held that spanned the entire spectrum of possible subjects.

I thank Dr. David Cheishvili, for being a true friend, even though he always reminded me that he was old enough to be my father. David was certainly the greatest source of both entertainment and scientific discussion during my time in the laboratory and, underneath his misleading shell, is one of the most caring people that I know. I am grateful to him for always making me feel better, nearly every day, by so obviously and so consistently losing to me in everything we competed in, which, over more than seven years, shifted from intense rubber band fights to Clash Royale to online billiards and finally to chess. I also thank Dr. Lisa Bureau, who probably doesn't remember me, but who spent my first two weeks as an undergraduate researcher teaching me the basics of molecular cloning (before I briefly left the lab for New York and returned to find her gone) and thus formed the foundation for all the experiments that I have done since. I thank Sergiy Dymov and Dr. Farida Vaisheva, my fellow lab mates for all these years, who tolerated me and laughed with me all this time. I thank Dr. Lax, who was, during his time as a postdoc in the lab, a great resource for scientific discussions. I thank Steffan Christiansen, who was a master's student during my first year of undergraduate work as we fought to

arrive first, near 9am, and leave last, past midnight, to impress our new boss, and spent the many hours in between helping each other. I thank our other past lab members who contributed their time and interest to me and my projects: Dr. Cristiane Matte, Dr. Yan Wang, Dr. Zsofia Nemoda, and Chengjian “Jim” Cao. Lastly in our lab, I thank all of the many undergraduate students who, beyond being of great assistance, in letting me teach them, taught me how to teach: Xinyu Zhao, Tatsuya Corlett, Tara Shomali, Emma Paulus, Peter Jeon, Claire Lin, Mor Robin, Claire Butler, and David Gao. Outside of our lab, I thank the collaborators that let me contribute to their projects, Dr. Gal Yadid and Dr. Ehab Abouheif.

There are no words to express the deepness of my gratitude to my fellow graduate students and my friends – most particularly Dr. Anne-Sophie Pepin and (soon-to-be Dr.) Andrew N. Bayne – for their closeness and support throughout the years. Thank you for listening to me complain about and helping me to solve every problem I encountered, as I tried to do for you, and all the fun alongside and outside the realm of science. An equal portion of my gratitude goes to my older friends – Ben Dizik, Alex Rirak, Dan Ragolsky, Daniel and Nathan Aleynick, and Muny Sidhu – who have so constantly supported me all this time, despite the distance. I thank my girlfriend, Jessica Tabarah, for her encouragement, her joy, her care, and her tolerance for those late nights and weekends in the lab. I am grateful also for the help of my committee members, Dr. Daniel Bernard, Dr. Paul Clarke, Dr. Ehab Abouheif, and Dr. David Dankort, for their guidance and support, in formal and informal settings, as well as that of professors Dr. Terry Hebert and Dr. Jean-François Trempe. Importantly, I thank our wonderful administrative staff and particularly Tina Tremblay, who worked so hard to ensure the success and happiness of all Pharmacology & Therapeutics graduate students. Finally, I thank the technicians and staff who have helped me with my experiments along the way: Dr. Nicolas Audet, Wolfgang Reintsch, Yaned Gaitan, and Julien Leconte. It is also worth noting that acknowledgements specific to each project can be found in the acknowledgements section of each chapter that makes up the body of this thesis.

6.1 Sources of funding

The majority of this work was sponsored by the Canadian Institutes for Health Research (PJT159583). Daniel M. Sapozhnikov and his work also received funding from the Natural Sciences and Engineering Research Council (NSERC) for other projects not included in this thesis (RGPIN-2016-04792). Daniel M. Sapozhnikov also received the following financial support from the McGill University Faculty of Medicine and Health Sciences: the J.P Collip Fellowship, the Friends of McGill Fellowship, and two differential fee waivers. Daniel M. Sapozhnikov also received the following financial support from the McGill University Department of Pharmacology and Therapeutics: GREAT Travel Award and Graduate Excellence Award. Daniel M. Sapozhnikov also received two travel awards from the Canadian Epigenetics, Environment and Health Research Consortium and Keystone Symposia.

Contributions of Authors

7.1 Chapter 1: Introduction

The doctoral candidate wrote this chapter. Sections 11.2.3-11.2.5 of Chapter 1 contain portions of text and figures that were published in *Biomedicines*[1] and were written by the doctoral candidate and edited by Dr. Moshe Szyf.

7.2 Chapter 2: Unraveling the functional role of DNA demethylation at specific promoters by targeted steric blockage of DNA methyltransferase with CRISPR/dCas9

The doctoral candidate wrote the text, designed all experiments, performed all experiments, and performed all analyses. Dr. Moshe Szyf guided the research, assisted in experimental design, and edited the text.

7.3 Chapter 3: Bisulfite-free sequencing of oxidized cytosines (APOBEC-seq) reveals a causal role of TDG in oxidized promoter re-activation and its ubiquitous MBD3-mediated presence in active gene promoters

The doctoral candidate wrote the text, designed all experiments, performed all experiments, and performed all analyses. Dr. Moshe Szyf guided the research, assisted in experimental design, and edited the text.

7.4 Chapter 4: General Discussion

The doctoral candidate wrote this chapter. Portions of this text were published in *Biomedicines*[1] and *Nature Protocols*[3] and were written by the doctoral candidate and edited by Dr. Moshe Szyf.

List of abbreviations:

5caC: 5-carboxylcytosine

5fC: 5-formylcytosine

5hmC: 5-hydroxylcytosine

5mC: 5-methylcytosine

AAV: adeno-associated virus

AP site: apurinic/apyrimidinic site

APEX1: apurinic/apyrimidinic endodeoxyribonuclease 1

APOBEC: apolipoprotein B mRNA-editing enzyme, catalytic polypeptide

BAM: binary sequence alignment map

BER: base excision repair

BFP: blue fluorescent protein

bp: base pairs

C5: carbon 5 of the pyrimidine ring of cytidine

C6: carbon 6 of the pyrimidine ring of cytidine

CAB-seq: chemical-assisted bisulfite sequencing

caCLEAR: 5caC clearance

CAR: chimeric antigen receptor

Cas9: CRISPR associated protein 9

CAT: chloramphenicol acetyltransferase

CG/CpG: cytosine followed by guanine in the DNA sequence

CGI: CpG island

ChIP: chromatin immunoprecipitation

ChIP-qPCR: chromatin immunoprecipitation followed by quantitative polymerase chain reaction

ChIP-seq: chromatin immunoprecipitation followed by sequencing

CMV: cytomegalovirus

CRISPR: clustered regularly-interspaced short palindromic repeats

dCas9: nuclease-dead CRISPR associated protein 9

DNA: deoxyribonucleic acid

DNMT: DNA methyltransferase

DRB: 5,6-Dichloro-1-beta-D-ribofuranosylbenzimidazole

EDTA: ethylenediaminetetraacetic acid

eGFP: enhanced green fluorescent protein

EMSA: electrophoretic mobility shift assay

EM-seq: enzymatic methyl-sequencing

eRNAs: enhancer RNAs

ESCs: embryonic stem cells

FACS: fluorescence-activated cell sorting

FBS: fetal bovine serum

GEO: Gene Expression Omnibus

GFP: green fluorescent protein

gRNA: guide RNA

gRNAscr: scrambled non-targeting guide RNA

H3K27ac: histone 3 acetylated on lysine 27

H3K27me3: histone 3 tri-methylated on lysine 27

H3K4me1: histone 3 methylated on lysine 4

H3K4me3: histone 3 tri-methylated on lysine 4

H3K9ac: histone 3 acetylated on lysine 9

H3K9me3: histone 3 tri-methylated on lysine 9

HDAC: histone deacetylase

HRP: horseradish peroxidase

IgG: immunoglobulin G

INDEL: insertion or deletion

IP: immunoprecipitation

KO: knockout

LC-MS/MS: liquid chromatography tandem mass spectrometry

LPS: lipopolysaccharide

MAB-seq: methylase-assisted bisulfite sequencing

MBD: methyl-binding domain or methylated DNA binding domain

mESCs: mouse embryonic stem cells

NEB: New England Biolabs

NTP: nucleoside triphosphate

NuRD: Nucleosome Remodeling Deacetylase

OGT: O-GlcNAc transferase

ONT: Oxford Nanopore Technologies

PacBio: Pacific Biosciences

PAM: protospacer adjacent motif

PBS: phosphate-buffered saline

PCR: polymerase chain reaction

PIC: protein inhibitor cocktail

pol2-PS5: RNA polymerase II phosphorylated on serine 5 of the C-terminal domain

poly(I:C): polyinosinic:polycytidylic acid

PRC2: polycomb repressive complex 2

qPCR: quantitative polymerase chain reaction

rDNA: ribosomal DNA

RNA: ribonucleic acid

RNA-polII/polII/pol2: RNA polymerase II

rpm: rotations per minute

rRNA: ribosomal RNA

RSB: reticulocyte standard buffer

RT-qPCR: reverse transcription followed by quantitative polymerase chain reaction

SAgRNA: S. aureus guide RNA

SAH: S-adenosyl homocysteine

SAM: S-adenosyl methionine

SDS: sodium dodecyl sulfate

SMRT: single-molecule, real-time

SNP: single nucleotide polymorphism

SPgRNA: S. pyogenes guide RNA

SRA: SET and RING finger-associated

SV40: simian virus 40

T4-BGT: beta-glucosyltransferase from T4 phage

TBP: TATA-box binding protein

TBST: Tris-buffered saline with Tween-20

TDG: thymine DNA glycosylase

TDGKO: thymine DNA glycosylase knockout

TES: transcription end site

TET: ten-eleven translocation

tRNA: transfer RNA

TSA: trichostatin A

TSS: transcription start site

UTR: untranslated region

WGBS: whole-genome bisulfite sequencing

ZBTB: BTB/POZ family of zinc-finger (ZF) proteins

ZF: zinc finger

List of figures:

Chapter 1: Figures

Chapter 2: Figures

Chapter 2: Supplementary Figures

Chapter 3: Figures

Chapter 3: Supplementary Figures

List of tables:

Chapter 2: Supplementary Tables

Chapter 3: Supplementary Tables

Chapter 1: Introduction

11.1 The role of DNA methylation in gene expression

The term "DNA methylation" broadly refers to the covalent attachment of a methyl (CH_3) group to nucleobases that make up deoxyribonucleic acids (DNA). However, methylation of only cytosine and adenine have been reported to occur naturally[4]. Adenine methylation (N^6-methyladenosine) is the most prevalent DNA modification in prokaryotes[5] and has impacts on gene expression regulation, DNA damage repair, cell cycle regulation, and in the immune response[6] of which, perhaps, the simplest example is the methylation of the internal adenine of EcoRI sites (those comprising of the sequence GAATTC) in the *E. coli* genome, thus preventing their cleavage by the EcoRI restriction enzyme, while invading foreign DNA is unmethylated at these sites and can be efficiently cleaved by EcoRI and is therein degraded[7]. However, the presence of N^6-methyladenosine in eukaryotes remains a topic of debate; although it was first reported to be detected in eukaryotes in 2015[8,9] and in mammals in 2016[10], its extremely low frequency and the resulting potential for artifactual contamination have led to numerous contradictory reports[11-14].

Cytosine methylation (5-methylcytosine, 5mC), on the other hand, is found at high frequencies in eukaryotes – including in mammals – and in plants, fungi, as well as in bacteria, in which its functions mimic those of N^6-methyladenosine[15]. Cytosine methylation is most abundant in plants, with 5mCs accounting for as many as 50% of all cytosines[16]. In humans, 5mCs account for only approximately 5% of all cytosines[17], but because mammalian cytosine methylation occurs primarily in the context of CG dinucleotides (CGs or CpGs), CpG methylation rates are similarly high, ranging from 60-90%[18]. It is interesting to note that while DNA methylation is critical to human and mouse development[19], it is practically absent in *Drosophila*, most yeast strains, and *Caenorhabditis elegans*. N^4-methylcytosine, also prevalent in bacteria, has recently been reported to be detected in simple eukaryotes[20], but not in mammals. Due to its

high abundance, CpG methylation is the most well-studied modification: this thesis focuses on mammalian CpG methylation due to the large body of supporting scientific evidence and due to its implications in human health and thus, herein the term “DNA methylation” will refer exclusively to CpG methylation.

11.1.1 5-methylcytosine detection methods

5mC was first discovered in 1948 by Rollin Hotchkiss from a preparation of calf thymus subjected to paper chromatography[21]. The movement of a boundary of n-butyl alcohol along filter paper successfully separated the four bases of DNA, but unexpectedly yielded an apparent fifth base migrating faster than cytosine, which he correctly proposed to be 5-methylcytosine. While critical for the discovery of 5mC, paper chromatography reveals no insight as to where in the DNA the 5mC is present. Accurate detection and quantification of specific genomic locations of DNA methylation is an essential prerequisite for understanding its functional implications in basic molecular biology and gene expression regulation and its roles in development, disease, and environmental responses. Broadly, such sequence-aware detection methodologies can be separated into several categories that are methodologically distinct but share some technical overlap: (1) restriction enzyme based approaches, (2) affinity based approaches, (3) polymerase chain reaction (PCR) based detection, and (4) sequencing based technologies.

Historically, the simplest analysis of DNA methylation invoked the natural intolerance of specific bacterial restriction enzymes to the presence of 5-methylcytosine in their target sites[22]. A classical example is the HpaII/MspI restriction enzyme combination wherein both enzymes cleave the CCGG sequence in DNA, but HpaII is inhibited by methylation of the internal cytosine[23]. By treating DNA of interest with each enzyme separately and thereafter amplifying by PCR a region of interest that contains at least one HpaII/MspI site, it is possible to estimate by gel electrophoresis the amount of inhibition of HpaII digestion as compared to that of MspI. This produces a semi-quantitative profile of CG

methylation in a region of interest, but lacks both the higher quantitative accuracy and – in the case of multiple HpaII/MspI sites within a PCR amplicon – also the single CG resolution and the possibility for genome-wide interrogation which were introduced in newer technologies.

An alternative set of techniques for DNA methylation quantification rely on the affinity capture of methylated DNA. This can be achieved with either a 5mC-specific antibody (MeDIP)[24] or a protein domain that exhibits preferential binding to 5mC-containing DNA (MethylCap[25] or MBD-seq[26]). DNA captured by both these methods can be quantified by quantitative PCR (qPCR) or sequenced with next-generation sequencing technologies to compare across samples and again produce a semi-quantitative readout of DNA methylation levels. The major drawback of these approaches is that efficient capture of methylated DNA typically requires high methylation density (multiple sites) while efficient differential capture may necessitate large changes in DNA methylation, often rendering these methods unsuitable for accurate DNA methylation quantification and for single CG level analyses.

Newer technologies largely depend on an initial step consisting of what is widely considered to be the “gold standard” protocol for DNA methylation analysis: the conversion of DNA with sodium bisulfite[27]. Bisulfite conversion, as it is typically referred to, actually consists of several chemical steps, wherein addition of sodium bisulfite to denatured DNA results in a nucleophilic substitution reaction in which the bisulfite ion (HSO_3^-) replaces the amino group ($–NH_2$) of cytosine. Then, in a desulphonation step, alkaline treatment results in the dissociation of the bisulfite group in favor of an oxygen atom, thus achieving a complete cytosine to uracil conversion[28]. Though this chemical conversion was reported in the early 1970s, it was not until 1992 that it became clear that the presence of a methyl group at the 5th carbon of cytosine (i.e., 5mC) prevented the nucleophilic substitution in the sodium bisulfite reaction, meaning that 5mC was nonreactive to bisulfite conversion[27] and thus, after an efficient bisulfite conversion

reaction, 5mCs would remain cytosines while unmethylated cytosines would become uracils. In a following PCR amplification step, uracils, which are not naturally found in DNA, can then be replaced with thymines in the amplified DNA. This produces a simple binary readout of DNA methylation at every cytosine on every strand of DNA (e.g., by sequencing), simultaneously engendering the possibility of detecting single molecule DNA methylation profiles as well as, given a large number of DNA copies (i.e., from a tissue or cell line), highly accurate DNA methylation quantification counted as C:T ratio at each position in the population of DNA strands. Unfortunately, this harsh chemical treatment can result in the degradation of as much as 96% of the DNA[29], rendering it less suitable for studies where the input DNA quantity might be limited. Therefore, in recent years, the field has moved towards an analogous bisulfite-free technique – called enzymatic methyl sequencing (EM-seq) and developed by New England Biolabs[30] – which boasts increased DNA quality after the conversion process. Here, TET2 and T4 phage β-glucosyltransferase (T4-BGT) are used to protect 5mC from a subsequent deamination reaction which invokes APOBEC3A to deaminate unmethylated cytosines to become uracils and results in a binary readout of DNA methylation equivalent to that of bisulfite conversion. Both methods typically exhibit unmethylated cytosine conversion rates upwards of 99%[30,31].

The least sensitive and seldom used detection techniques used after bisulfite conversion are methylation-specific PCR (MSP)[32] and methylation-specific high-resolution melting (MS-HRM)[33]. MSP invokes two different primer sets for the same region of interest, a methylated primer set that presumes all CG cytosines will remain cytosines after conversion and an unmethylated primer set that assumes all CG cytosines will be converted to uracils. Comparing the amplification of the two primer sets – optimally, by qPCR – produces a semi-quantitative profile of DNA methylation, but it is not high-throughput, subject to amplification and primer design biases and efficiencies, limited in CG coverage, and not sufficiently quantitative. MS-HRM, on the other hand, uses a single primer set and relies on the higher bond strength of C – G base pairs (i.e.,

of methylated cytosines after conversion) as compared to T – A base pairs (i.e., of unmethylated cytosines after conversion) which can be observed as differential melting curves following qPCR. This method has numerous disadvantages as it provides very limited quantitative information, is sensitive to several biases involved in PCR, and seldom produces information at the single CG level.

Taking these drawbacks into consideration, the single most accurate DNA methylation detection method after bisulfite conversion is by DNA sequencing[34]. As mentioned previously, sequencing can be applied to single DNA strands or large populations of DNA, as well as to single targeted regions or genome-wide sequencing, making sequencing highly accurate, sensitive, robust, and flexible to experimental design and budget considerations. Single amplicons can be subcloned directly from PCR reactions into commercially available TA vectors and sequenced by Sanger sequencing to obtain allelic patterns of methylation if such research questions are of interest[35]. Separately, to quantify the DNA methylation level (in a large number of DNA copies from a sample of interest) of a single region or small set of DNA regions, this limited number of regions can be amplified with specifically designed primers and sequenced at extremely high coverage with several targeted next-generation sequencing technologies, such as pyrosequencing[36] or Illumina's MiSeq[37] technology, to obtain high read depth and therefore determine with high accuracy the single CG level DNA methylation levels. Naturally, this can be extended using shotgun library preparation methods to perform whole-genome bisulfite sequencing (WGBS)[38] or EM-seq to, in a hypothesis-free manner, discover any regions in the genome which might exhibit DNA methylation changes from an experimental, physiological, or pathological condition. WGBS has been successfully applied across hundreds of studies to identify single CGs or regions which display differential DNA methylation under such circumstances.

It also interesting to note the more recent ability to directly determine DNA methylation levels in single strands of DNA – without the need for any conversion – by using either

of two long-read sequencing technologies: single molecule real-time (SMRT) sequencing from Pacific Biosciences (PacBio) and nanopore sequencing from Oxford Nanopore Technologies (ONT)[39]. These technologies detect 5mC in the raw sequencing signal and, though they avoid biases introduced by conversion, are limited by the fact that DNA cannot be amplified by PCR as it would erase 5mC in the amplified DNA and thus they demand relatively high input DNA quantities that prohibit experiments involving single cells, small organisms, or other so-called precious (limited in quantity) tissue samples.

Together, these technologies have allowed researchers to develop a thorough fundamental understanding of DNA methylation dynamics and explore its diverse contributions to biology.

11.1.2 Enzymes which directly modify DNA methylation levels

Methylated cytidine triphosphate is not a naturally occurring nucleoside triphosphate (NTP) and thus 5mC is not incorporated into the DNA as it is being extended by DNA polymerase. Instead, 5mC is formed by the catalytic action of DNA methyltransferase enzymes on unmethylated cytosine. Humans have three such enzymes encoded by three separate genes: DNMT1, DNMT3A, and DNMT3B. All three enzymes rely on the methyl donor S-adenosyl-methionine (SAM) to form 5mC and produce S-adenosyl-homocysteine (SAH) as a by-product. Knockout of any of the three DNMTs is lethal in mice[40,41].

DNMT1 is unique in that it acts primarily as a maintenance methyltransferase by copying parental strand DNA methylation patterns to newly synthesized daughter strands, thereby ensuring the fidelity of the genome-wide DNA methylation pattern through cell division and DNA replication. Mechanistically, this behavior stems from both the preference of DNMT1 to methylate the unmethylated CG in hemi-methylated double-stranded DNA (i.e., methylated CG opposite an unmethylated CG in the

complementary strand), rather than double-stranded DNA in which both CGs are unmethylated, and from the recruitment of DNMT1 to the replication fork in actively replicating DNA, which contains such hemi-methylated DNA in the form of a methylated parental DNA strand in complex with a newly synthesized unmethylated daughter strand. DNMT1 is also considered to be highly processive, requiring no energy for diffusion along a DNA strand[42] and thereby ensuring faithful maintenance methylation. This maintenance activity is not without exceptions, as DNMT1 has been reported to display some “de novo” methylation activity of wholly unmethylated DNA, particularly when it is made to be single-stranded, *in vitro*[43] and *in vivo*[44], though with a dramatically reduced processivity[42]. The recruitment of DNMT1 to the replication fork has been repeatedly shown to be dependent primarily on an interaction with PCNA[45,46], a DNA clamp critical for processivity of DNA polymerase delta at the replication fork, as well as with UHRF1, which also binds hemi-methylated DNA. Though classically depicted as having no sequence specificity beyond the CG dinucleotide[42], recent evidence suggests that DNMT1 may indeed have a preference for the sequence composition that flanks the CG site[47], thereby raising the interesting possibility that the genomic sequence inherently governs the establishment of the genomic methylation pattern to a greater extent than previously thought[48].

Unlike DNMT1, DNMT3A and DNMT3B are largely considered to be “de novo” methyltransferases[41] due to their lack of requirement of hemi-methylated DNA as a template for methylation activity. However, numerous studies have suggested that DNMT3A and DNMT3B nonetheless contribute to maintenance methylation[49-51]. Still, the major role of these enzymes is in the deposition of novel methyl marks in response to internal and external stimuli. Their expression is highest in undifferentiated cells during embryonic development[41,52], where their activity is necessary for survival and leads to highly tissue-specific patterns of DNA methylation in differentiated cells[53]. DNMT3A and DNMT3B also are the executors of the dynamic component of the DNA methylation landscape as they mediate the establishment of new methylation patterns in

adults in response to diverse stimuli, such as immune stimulation[54], addiction and learning[55], cancers[56], and post-traumatic stress disorder[57]. Like DNMT1, DNMT3A and DNMT3B exhibit some sequence specificity for the bases that flank the CG dinucleotide, but this can be reduced by their recruitment to nonoptimal sequences by other protein factors, such as the highly related but non-catalytic DNMT family member, DNMT3L[58], which also increases their methylation activity[59]. In fact, their sequence specificities appear to contribute to the small degree of non-CpG methylation in the human genome[60,61], which DNMT1 is believed to be incapable of[61].

The complex and dynamic CpG landscape is likewise sculpted by demethylation. The simplest mechanism for demethylation is via the passive route[62]: it is lost in dividing cells during the synthesis of new DNA in the absence of maintenance methyltransferase activity. Yet, this cannot explain all of the demethylation observed in humans, such as the regional demethylation in nondividing postmitotic neurons. In fact, an active form of demethylation is required for learning and memory[63,64], as is active methylation[50,65,66]. Together, passive and, to a lesser degree, active demethylation are responsible for two demethylation waves during development[62,67]. The active DNA demethylation process is initiated by the ten-eleven translocation family of proteins (TET1-3), which directly modify the 5mC moiety by several consecutive oxidation reactions, yielding 5-hydroxymethylcytosine (5hmC), 5-formylcytosine (5fC), and, ultimately, 5-carboxyl-cytosine (5caC)[68]. The implications of this oxidation activity on the active DNA demethylation pathway are described in a later section (11.1.8 Active DNA demethylation pathway). TET proteins are naturally sensitive to environmental factors as their catalytic function requires iron, α-ketoglutarate, and ascorbic acid, they interact with dozens of nuclear proteins[68], and they are modified by numerous posttranslational modifications[68]. The three proteins appear to have nonredundant functions, as TET3 knockout is lethal, while TET1 or TET2 knockout yield developmentally normal animals, but combined TET1/TET2 knockout is also often lethal[69]. Apart from embryonic development, TET proteins contribute to dynamic changes in DNA methylation in

hematopoiesis, cancers, immune responses, and learning and memory, as well as other adult processes[69]. The ability of other human proteins to catalyze active DNA demethylation – such as, perhaps, that of MBD2[70] – remains unclear[71-73] and, thus, TET-mediated active DNA demethylation remains the only widely accepted mechanism of active DNA demethylation in mammals.

11.1.3 Proteins which interact with methylated DNA

One of the ways in which 5mC affects cellular phenotypes is by the altered recognition of other DNA-binding proteins to their target sequences when these targets are methylated or unmethylated. The differential ability of proteins to bind their targets as a function of methylation is well defined across a large catalogue of methylation-sensitive bacterial restriction enzymes and includes both enzymes inhibited by methylation of their target sites and those that only cleave methylated sequences[74]. There is also considerable evidence of methylation-sensitive transcription factors in humans. The binding of most major classes of transcription factors is directly inhibited by the presence of 5mC in their binding site; however, numerous transcription factors preferentially interact with 5mC[75].

The most well-studied example of proteins that specifically interact with 5mC is the highly conserved methyl-binding domain (MBD) family of proteins (MBD1-4 and MECP2)[76]. Initially identified in MECP2[77], the MBD confers selective recognition of 5mC and therefore forms the basis for the aforementioned affinity-based purification of methylated DNA. Unlike the other MBD family members, MBD3 lacks four conserved amino acids in the MBD and is widely believed to not bind methylated DNA[78], though this hypothesis has been recently challenged[79]. Interestingly, knockout of only MBD3 results in embryonic lethality[76], emphasizing its as-of-yet mechanistically elusive importance to embryonic development despite an evolutionary loss or reduction of methyl-CpG binding capacity. The MBD proteins also display sequence specificity for sites flanking the CG dinucleotide, with varying preferences and degrees[80,81]. The

recruitment of MBD family members to methylated DNA is largely considered to result in gene silencing by interference with the binding of other transcription factors and by other repressive activities of these proteins and the complexes in which they reside[76], elaborated upon further in a subsequent section (11.1.7 Codependence of DNA methylation and other epigenetic marks). Here, an exception is MBD4, which exhibits an enzymatic N-glycosylase activity and likely functions in DNA repair[76].

Other proteins may also directly recognize 5mC. As mentioned above, UHRF1 is a critical sensor of hemi-methylated DNA required for high-fidelity maintenance methylation by DNMT1. The 5mC interaction of UHRF1 is mediated by its SET and RING finger-associated (SRA) domain[82], which is also found in UHRF2, but the latter protein has little binding affinity towards 5mC[83]. UHRF1 is also required for embryonic development, unlike UHRF2, which tends to be expressed at higher levels in differentiated adult tissues[80]. Another well-established group of 5mC-binding proteins consists of Kaiso and Kaiso-like proteins[80]. Kaiso is a member of the BTB/POZ family of zinc-finger (ZF) proteins (ZBTB) and both Kaiso and two related proteins, ZBTB4 and ZBTB38 preferentially binding methylated cytosines, though Kaiso requires methylated CG doubles (CGCG) for high affinity binding whereas ZBTB4 and ZBTB38 can bind single methylated CpGs[84,85]. All three ZBTB proteins are believed to contribute to transcriptional repression upon 5mC binding[85,86]. Recent advances have allowed larger scale screening of transcription factors for methylation sensitivity[75,87]. These studies have revealed a rather large number of proteins that preferentially bind 5mC and include key developmental genes, such as HOXC11, HOXB13, POU5F1/OCT4, and NKX proteins, suggesting a potentially broad and interrelated network of proteins that act as the effectors of the 5mC signal.

11.1.4 The genomic landscape of DNA methylation

The DNA methylation patterns in humans are so inherently tissue-specific that DNA methylation is more similar in the same tissue across individuals than across the

different tissues of one individual[88] and, moreover, differential DNA methylation patterns can be used to identify cell types within cell mixtures[89]. Yet, the corollary of this phenomenon is that both inter-individual tissue-specific DNA methylation similarities and the fraction of invariable methylated or unmethylated CpGs represent certain ubiquitous properties of the DNA methylation profile. To better understand this characteristic genome-wide DNA methylation pattern, it is critical to bear in mind that CpGs are statistically depleted from the genome, occurring at approximately one-fifth of the expected rate based on genomic GC content[90]. This is widely believed to stem from the elevated mutagenic potential of 5mC, which is prone to spontaneous deamination to become thymine at a rate that is 2.0 to 3.2 times higher than that of unmethylated cytosine[91]. The reduced rate of unmethylated cytosine mutation, in turn, likely stems from the deamination-induced formation of uracil, an unnatural DNA base that triggers DNA repair and more frequent correction of the mutation[91]. However, it is also possible that the dynamic functional regulation enabled by the binary methylation status of CpGs has also driven the evolution of a CpG-scarce human genome.

The implication of this underrepresentation is that CpGs occur at a low density throughout the entire genome, except for specific regions of high CpG frequency, termed CpG islands (CGIs) and containing approximately 5% of all CpGs[92]. CGIs are typically strictly defined as regions of at least 200 base pairs (bp) in length characterized by a GC content above 50% and an observed to expected CpG ratio greater than 60%[93]. Despite the fact that most genomic CpGs are methylated[18], CpGs in CGIs are typically uniformly unmethylated, with genome-wide methylation estimates of only 6-8%[94]. Interestingly, approximately 50% of CGIs occur within gene promoters[95] and, reciprocally, 72% of gene promoters contain a CGI[96]. Genes with CGI-promoters are more frequently constitutively expressed and represent so-called “housekeeping” genes[96] and 90% of all such genes are estimated to contain CGI-promoters. Thus, unmethylated CGIs represent an important evolutionarily conserved and regulatory aspect that promotes gene expression in the human genome. Yet, the small fraction of

genes with CGIs that can be methylated conversely represent a class of genes with tissue-specific gene expression[97-99], while aberrantly hypermethylated CGIs are a hallmark of nearly every cancer type[100], exemplifying the functional relevance of differential DNA methylation of CGIs to cellular identity.

While CGIs contain the most ubiquitously unmethylated CpGs, constitutively methylated regions are also extensively observed among the remaining 95% of CpGs scattered across the human genome. It is estimated that over 90% of these methylated CpGs occur in repetitive elements[101], which account for 50-70% of the genome and vastly outnumber the 1-2%[102] of DNA bases which are part of protein-coding genes. Repetitive elements are a diverse class of DNA sequences which exist as multiple copies in the genome and have been attributed both positive contributions critical to basic cellular characteristics – such as replication – as well as an accompanying spectrum of negative effects[103]. The dangers posed by repetitive elements are numerous and stem from the ability of many of these transposable-element-derived sequences to move throughout the genome and disrupt critical genetic elements, their potential to drive inter- and intra-chromosomal recombination events on the basis of their sequence similarity, their tendency to form secondary structures that interfere with the replication machinery, and various other routes that lead to genomic instability and jeopardize cellular health[103]. Therefore, numerous regulatory mechanisms have evolved in humans to minimize the risk posed by repetitive sequences, including their consistent heterochromatization[103], which refers to highly compacted DNA believed to primarily function to repress the activities of such repetitive elements by restricting their physical accessibility and is characterized by distinct histone modifications accompanied by extensive DNA methylation[104]. DNA methylation is thought to play a major role in the formation and persistence of heterochromatin[104] (detailed in Section 11.1.7) within repetitive elements and without and loss of DNA methylation in repetitive elements has been observed in numerous hereditary disorders[103] – such as systemic lupus erythematosus[105] – as well

as in neurological disorders[106,107], most cancers[103,108], aging[109], and many other pathologies.

Another major reservoir of methylated CpGs in the human genome is in the bodies of genes[110]. The first exon, which is near to the promoter, typically remains unmethylated in expressed genes while, in contrast, gene body methylation is a consistent feature of expressed genes across eukaryotes[111]. The function of this methylation is discussed in Section 11.1.6. Though gene bodies are typically CpG-poor[110], they also represent a major exception to the observation that CGIs are typically unmethylated: CGIs occurring in gene bodies, as well as those generally outside of gene promoters, often can be methylated, particularly in a tissue-specific manner[112]. For example, it is estimated that as many as 34% of intragenic CGIs in the brain are methylated[113]. CGIs outside of promoters often act as enhancers[114], which regulate local and distal gene expression in a tissue-specific manner[115].

Finally, in a manner similar to intra- and intergenic CGIs, CpG-poor promoters exhibit some of the most dynamic DNA methylation patterns. In general, CpG-poor promoters are more likely to be expressed in a tissue-specific manner and the CpGs within them are typically methylated in tissues where the gene is not expressed[116,117].

11.1.5 The function of DNA methylation in gene promoters

Despite the fact that the discovery of methylated cytosine in 1948 precedes the discovery of the DNA double helix, it was not until the late 1970s and early 1980s that the function of methylated cytosine began to be elucidated. Initial evidence for a role of 5mC in gene expression regulation was correlational: several groups observed that, in non-human organisms, gene promoters were more active in tissues in which they were unmethylated compared to tissues in which the same promoter was methylated[118,119]. Similar observations – including those in humans – continued to be published over several years and formed the basis of an accepted but as of yet speculative hypothesis

that DNA methylation of gene promoters might lead to the downregulation of gene expression[15]. Yet, it remained unclear if DNA methylation of promoters directly caused gene downregulation or if it was simply a marker of inactive gene promoters. It was not until 1983 – fostered by advances in cloning technologies – that this hypothesis was truly tested. In a landmark article[120], Kruczek and Doerfler demonstrated that in vitro methylation (by bacterial HpaII or HhaI methyltransferases) of a plasmid expressing a chloramphenicol acetyltransferase (CAT) reporter driven by an adenoviral E1a promoter silenced reporter activity upon transfection into mouse cells. They attributed this affect to the methylation of two HpaII (CCGG) or three HhaI (GCGC) target sites in the core E1a promoter and observed further that two alternative promoters were insensitive to methylation: one lacked these CG-containing sites completely and the other bore sites that were far (greater than 300 bp) from the recently discovered TATA box (note that the TATA box is a genetic element that binds TBP, which is critical for RNA polymerase recruitment and ensuing gene expression[121]). Kruczek and Doerfler also reported that methylation of the coding sequence of CAT did not affect expression. Together, these careful experiments provided the first causational evidence of the silencing effect of DNA methylation and informed future research of the variable impact of DNA methylation as a function of proximity to key regulatory elements in the promoter.

Remarkably, these fundamental descriptions of promoter methylation by Kruczek and Doerfler remain wholly valid today; still, decades of DNA methylation research have advanced our understanding of the molecular mechanisms by which promoter methylation silences gene activity. These mechanisms can be separated into two main categories: (1) the direct inhibition of transcription factor binding to DNA and, (2) indirect transcriptional repression mediated by proteins which recognize 5mC.

Perhaps the simplest mechanism by which DNA methylation inhibits gene expression is by reducing the affinity of transcription factors to their binding sites in a manner that reflects the aforementioned inhibition of DNA cleavage activity of many bacterial

restriction enzymes by methylation of their target sites. Transcription factors recognize specific sequence motifs by an integrated base and shape readout[122]. Base readout refers to the interaction between the structural features or specific amino acids of transcription factors and the specific bases of the DNA, mediated by hydrophobic contacts, direct hydrogen bonds, water-mediated hydrogen bonds[122], and other interactions, such as the bidentate hydrogen bonds formed between arginine residues and guanine bases[123]. Transcription factors are also sensitive to DNA bending and unwinding states, forming the basis of a shape readout, which is also a function of the sequence of bases[122]. Together, these two properties define the binding profiles of sequence-specific transcription factors, such as the CG-binding activity of DNA methyltransferases and of the binding motif of SP1, which recognizes specific stretches of GC bases termed “GC boxes”[124]. Transcription factors, in turn, ultimately modulate the accessibility of the DNA and the recruitment and assembly of RNA polymerase complexes and thereby contribute to gene expression[125]. The addition of the hydrophobic methyl group to cytosines can directly affect DNA structure, widening the major groove and narrowing the minor groove of the double helix and thereby altering the shape readout of transcription factors[126]. The methyl group also alters the base readout of transcription factors as it modifies the landscape of potential interactions; it can allow hydrophobic contacts that are more reminiscent of a thymine base, which also has a 5-methyl group[126]. Both of these consequences of 5mC can reduce or even eliminate the ability of transcription factors to recognize and bind their target sequences and constitutes a common phenomenon across all major classes of transcription factors[75]. This is best examined from a causational standpoint by electrophoretic mobility shift assays (EMSAs), which can test the binding of recombinant proteins to the same DNA when it is methylated or unmethylated[127], and some well-known examples of transcription factor binding activity that is inhibited by DNA methylation as ascertained by these assays include that of TBP[128] and of the glucocorticoid receptor[129]. Thus, DNA methylation can directly reduce gene activity by interfering with the recruitment of positive regulators of gene expression in the promoter region.

The same effect of promoter DNA methylation on reduced gene expression can be mediated by the opposite activity of other proteins in specifically recognizing 5mC, examples of which were discussed previously in Section 11.1.3. The most well-established mechanism is through the recruitment of the MBD family of proteins to methylated promoters. The MBDs, in turn, primarily exert silencing effects through their recruitment of histone deacetylases (HDACs)[76]; the precise nature of these interactions is discussed further in Section 11.1.7. HDACs function to remove the acetyl group from histones, which are proteins around which the DNA is wound[130]. The positive charge of histone N-terminal tails underlies their interaction with the negatively charged phosphate backbone of DNA – serving to compact the DNA – while the enzymatic addition of an acetyl group neutralizes this positive charge, thus relaxing the histone-DNA interaction and increasing DNA accessibility to transcriptional machinery[131]. Acetylation also works to increase gene expression by the recruitment of transcription factors[131]. Therefore, the removal of the acetyl group by HDACs promotes tighter wrapping of the DNA around histones, reducing accessibility to transcriptional machinery and repressing gene expression[130]. Deacetylation can also lead to gene silencing by prohibiting the aforementioned transcription factor interactions and by allowing instead the deposition of other modifications to the histone tail (i.e., lysine methylation) which may foster the recruitment of other transcriptionally repressive proteins[130]. Despite this well-described 5mC-MBD-HDAC pathway to gene silencing by DNA methylation, a recent study reported that the combinatorial deletion of all MBDs did not lead to the re-activation of silenced genes[132], challenging this classical view and suggesting that, instead, direct inhibition of transcription factor binding by 5mC may be the main mode by which DNA methylation silences gene expression.

The regulation of gene activity by promoter methylation is the best-studied scenario in the DNA methylation research field. The presence of DNA methylation in silenced promoters is a widespread and fundamental aspect associated with tissue-specific gene

expression patterns. For example, the insulin gene is expressed almost exclusively in the beta cells of the pancreas and its promoter is unmethylated in these cells but methylated in most other tissues, including other neighboring pancreatic cell types[133]. Likewise, the promoter of the astrocyte marker, GFAP, is unmethylated in astrocytes while it remains methylated in surrounding neurons and other tissues[134]. Hepatocyte nuclear factor genes (e.g., HNF1A, HNF4A, etc.) too are transcriptionally active – with unmethylated promoters – predominantly in the hepatocytes of the liver[135]. Still, across these and other examples, it remains contested whether DNA methylation of the promoter is causal to gene silencing in the differentiated cells of other tissues, whether it was only causal at a key developmental stage, or if it is simply a marker of broader transcriptional processes that regulate expression. Nonetheless, promoter DNA methylation is a defining characteristic of tissue-specific gene expression.

Accordingly, aberrant hyper- or hypo-methylation of both CpG-poor and CGI gene promoters is commonly observed across most non-mendelian diseases and is thought to act as an alternative route to pathogenesis by the modulation of gene expression. In the same way that genetic mutation can lead to gene loss or, conversely, to hyperactivity to cause disease, promoter DNA hypermethylation can lead to gene expression loss while promoter DNA hypomethylation can increase gene activity. The etiology of cancer, for example, is classically attributed to, in part, the mutation and resulting functional loss of genes that normally inhibit or mediate the molecular pathways underlying the proliferative capacity, genomic integrity, or invasion potential of cells, commonly referred to as tumor suppressor genes: their genetic loss typically leads to a shift towards rapid cellular proliferation, mutation gain, or other properties that are critical to the development of cancer[136,137]. However, hypermethylation of the promoters of practically all tumor suppressor genes – often in the absence of genetic mutation – has been consistently observed across human cancers[138-142]. The same concept of hypermethylation-induced cancer risk extends to noncoding RNAs, which are also expressed from promoters and can have anticancer activities[143,144]. Conversely, the

hypomethylation and resulting increased expression of genes which function to positively regulate proliferative and other cancer-like capacities (i.e., oncogenes such as EGFR[145], RAS[146], and MYC[147]) is also widely reported in cancer. It is important to note that such DNA methylation changes – which thus appear to be able to contribute to cancer progression – are not necessarily directly causal and, while consequential, may sometimes instead be a result of other mutations that then lead to changes in methylation and gene expression. Such mutations may occur in any number of specific transcription factors or other nuclear factors or, more directly, in the components of the DNA methylation machinery, best exemplified by the strikingly high frequency of concurrent DNMT3A and TET2 mutation in lymphomas[148].

While cancer is an interesting and relatively straightforward example of the pathological consequences of both hyper- and hypo-methylation of gene promoters, the same principles apply to nearly all physiological and pathological states, which are tightly regulated by and dependent on gene expression levels. Bidirectional changes in the promoter methylation levels of the three aforementioned examples of tissue-specific genes are associated with distinct health states: (1) pancreatic insulin promoter hypermethylation in type 2 diabetes[149] and muscle cell insulin hypomethylation with exercise[150]; (2) astrocyte GFAP hypermethylation in glioma[151] and neural GFAP promoter hypermethylation in an animal model of autisim[152]; (3) colonic HNF4A promoter hypermethylation in necrotizing enterocolitis[153] and ovarian HNF1A promoter hypomethylation in clear cell carcinoma[154]. Examples of phenotype-associated changes in promoter DNA methylation extend well beyond these few cases and are not limited to exceptionally tissue-specific genes (particularly in cancer); evidence of a role for differential DNA methylation is commonly observed across the entire spectrum of health outcomes. Several other examples include glucocorticoid receptor promoter methylation in depression[155], adiponectin promoter methylation in insulin resistance[156], and altered alpha-synuclein promoter methylation in Parkinson's disease[157]. The breadth of such examples renders it clear that the link between gene expression silencing and DNA

methylation of gene promoters represents a broadly relevant and consequential attribute that is fundamental to human biology.

11.1.6 The function of DNA methylation in gene bodies

In contrast, the mechanisms by which gene body methylation regulates gene expression are comparatively less established. Gene bodies are generally CpG poor and contain repetitive elements[110]. Extensive methylation of gene bodies is a common feature of transcribed genes across all cell types[158,159] and gene body methylation typically correlates positively with gene expression levels[160,161], suggesting a potential role of gene body methylation in transcriptional regulation.

The primary function of gene body methylation is believed to be the silencing of alternative intragenic promoters by mechanisms that are not different from those by which DNA methylation silences expression at promoters. This function can be conceptually separated further into two distinct activities: the silencing of spurious transcription and the silencing of developmentally functional, true promoters[113]. Spurious transcription refers to the initiation of transcription from cryptic start sites which would then result in the production of noncanonical RNAs that do not represent conventional coding or noncoding products and, potentially, have a deleterious impact on cellular function due to unclear aberrant functions upon translation and the usage of cellular resources in their synthesis and, often, in their degradation by an RNA exosome complex in advance of translation[162]. DNA methylation may serve to mask such sequences from transcription factor binding and transcriptional initiation, thereby preserving the integrity of the transcriptome. On the other hand, intragenic regions capable of driving transcription may synthesize coding and noncoding RNAs that may be important for cellular function. Most human genes have well-described alternative promoters that result in protein isoforms that exhibit distinct behaviors[163]: it is possible that some fraction of intragenic promoters represent true, poorly hitherto described promoters that, for example, may have been unmethylated and functional during

developmental stages or have tissue-specific methylation and expression profiles[113], and thus the function of their methylation might mirror that of promoter methylation. Another somewhat semantic difference is that these intragenic promoter-like sequences may actually constitute an intragenic enhancer[164-166], rather than promoter, functionality of these elements. Though the main function of enhancers is to recruit transcriptional machinery and physically interact with distal promoters to increase the transcription output of the promoter, this means that there is significant functional overlap between the two classes of elements[167]. Like promoters, enhancers associate with RNA polymerase and therefore, in addition to their augmentation of production of gene products from promoters, can drive expression of the local sequence, typically referred to as enhancer RNAs (eRNAs), the quantity of which correlates positively to enhancer activity[167]. Aberrant intragenic enhancer activity and eRNA production has been shown to interfere with expression of the gene in which it is situated[164]; thus, DNA methylation may function to restrict transcription factor binding to enhancers when they are not active. Furthermore, enhancers by definition are expressed in a tissue-specific manner and are thought to functionally explain most tissue-specific gene expression[168]. Therefore, the methylation of gene bodies may in part reflect tissue-specific methylation patterns. Still, across all these scenarios, DNA methylation maintains a singular function in the silencing of transcription from intragenic regions. Accordingly, a loss of DNA methylation and an increase in spurious intragenic transcription is a hallmark of aging[169] and cancer[170].

Gene body methylation may also have additional roles in the regulation of transcriptional elongation and alternative splicing, which are inherently interrelated: variations in the elongation rate of RNA polymerase II have been shown to influence alternative splicing patterns, with slower elongation rates associated with increased inclusion of alternative exons in transcribed mRNA[171], the existence of distinct elongation checkpoints that regulate splicing[172], and a reciprocal effect of splicing factors on elongation rates[173]. DNA methylation of gene bodies appears to generally

reduce the rate of transcriptional elongation[174] and modify splicing patterns[175]. Though this is largely correlational evidence and may be a result transcription patterns which may in turn be sequence-dependent, there are several proposed mechanisms by which DNA methylation can regulate these processes. CTCF binds unmethylated DNA to, in part, interfere with transcriptional machinery and slow transcriptional elongation, promoting alternative exon inclusion[175]. The binding of CTCF is inhibited by methylation of its target site, which results in increased exon exclusion[176]. A second proposed mechanism is through the 5mC-mediated recruitment of MECP2, and though poorly defined, its effect on increased alternative exon inclusion is thought to be mediated by a sequence-specific preference for specific exons[177] and its interaction with splicing factors[177] and HDACs[178]. A final mechanism involves HP1, a protein which is recruited to methylated regions (though, indirectly, by recognizing other epigenetic marks typical of methylated regions) and also interacts with splicing factors[179]. Still, given that these contradictory mechanisms have not thus far been clearly delineated and, moreover, that DNA methylation appears to have little impact on the splicing of constitutive exons across which it is widespread[179], a role in alternative splicing appears to be an insufficient rationale to explain the consistent methylation observed in transcribed genes.

Another alternative view is that DNA methylation is ubiquitously deposited as a means for preserving genome integrity via its compaction and heterochromatization of DNA such that both the activity of widespread repetitive elements is suppressed and DNA is physically protected from damaging agents[180]. In this scenario, it is the unmethylated genome that is actively preserved in its unmethylated state, and therefore gene body methylation in many regions may be a by-product of these global forces. There is evidence for an active mechanism to preserve unmethylated DNA in the fact that TET enzymes[180] and other components of the active DNA methylation pathway (as demonstrated in the Chapter 3 of this thesis) are recruited to active genes. This would explain why the loss of DNA methylation at gene bodies, in some studies, leads to no

detectable increase in transcription initiation from intragenic sites[181]. In any case, the functional impact of gene body methylation still remains to be fully resolved.

11.1.7 Codependence of DNA methylation and other epigenetic marks

The tight relationship between DNA methylation and gene expression stems not just from the direct effects of DNA methylation on protein binding but also from the protein network that directly modifies and recognizes DNA methylation and is augmented by the vast number of protein interactions of that machinery and again by the functions and interactions of those proteins. Many of these proteins, in turn, work to establish and preserve DNA methylation patterns. Together, this “crosstalk” ensures the fidelity of the gene expression regulatory networks critical to cellular identity and survival. The entire gene expression regulatory network and the mechanisms that define its interdependence with DNA methylation is too vast a body of evidence to be wholly summarized herein, but there are several key interactions that form the foundation of this reciprocal relationship.

One definition of “epigenetics” is the collection of enduring changes to gene expression by modifications not related to the sequence of the DNA itself and involves changes mediated by DNA methylation and histone modifications, among other mechanisms. Histones interact with DNA as octamers composed of two copies each of H2A, H2B, H3, and H4[182] and there are at least nine different histone modifications (e.g., acetylation, methylation, phosphorylation, ubiquitylation, etc.) that can be covalently attached to hundreds of specific amino acids across these four proteins[182]. As mentioned above, the acetylation of histone tails near promoter regions is associated with actively expressed genes and is thought to be a functional consequence of both the relaxation of DNA and recruitment of interactors to the acetylated histone tail. Two commonly studied forms of histone acetylation are acetylation of lysine 27 on histone H3 (H3K27ac), predominantly found near active enhancers and promoters, and acetylation of lysine 9 on histone H3 (H3K9ac), which tends to be located near active

promoters[182]. On the other hand, histone methylation does not appear to have significant steric effects on DNA accessibility but exerts consequences on gene expression by recruiting specific protein interactors. Examples of well-studied modifications include tri-methylation of lysine 9 or 27 on histone H3 (H3K9me3 and H3K27me3), which are associated with gene silencing and heterochromatin, and tri-methylation of lysine 4 on histone H3 (H3K4me3), which is typically found in active promoters[182].

DNA methylation (and its absence) and specific histone methylations co-occur consistently throughout the genome; for example, unmethylated promoters and enhancers are enriched in H3K27ac[183] whereas silenced genes and heterochromatin typically exhibit DNA methylation and H3K9me3 or H3K27me3[104,184]. There are numerous dissected mechanisms by which DNA methylation changes lead to changes in histone modifications. For example, MECP2 both recognizes methylated DNA and interacts with the SUV39 protein family of H3K9 methyltransferases to deposit H3K9me3[185]. Likewise, MECP2 interacts with several HDACs through its transcriptional repression domain[186-188]. There is parallel evidence of interaction between other MBD family members and modifiers of histone modifications, such as the recruitment of HDACs to methylated DNA by MBD2[189] and the recruitment of the H3K9 methyltransferase SUV39H1 by MBD1[190].

However, the crosstalk between DNA methylation and histone modifications is not unidirectional. The aforementioned examples of the recruitment of histone modifiers by MBD family members can be interpreted oppositely, with histone modifications recruiting the MBD family members. Yet, the mechanisms of histone-modification-directed deposition of DNA methylation are perhaps more straightforward in the case of recruitment of the DNA methyltransferase enzymes, of which there are numerous examples. DNMT3A, DNMT3B, and DNMT3L all recognize unmodified H3K4 with their ADD domain, which appears to stimulate their activity and lead to local DNA

methylation[191,192]. The PWWP domains of DNMT3A and DNMT3B also recognize methylated H3K36, a mark typically found in gene bodies. The protein HP1, which recognizes methylated H3K9 through its chromodomain and interacts with MBDs[193], also recruits DNMT1 and DNMT3A and stimulates their activity[194,195]. The polycomb repressive complex 2 (PRC2), which possesses H3K27 methyltransferase activity, can also recruit DNMT3A, DNMT3B, and DNMT1[196]. Dozens of other interactions between histone modifiers or binders of modified histones and the DNA methylation machinery have been reported[192,197], creating a complex epigenetic network that regulates gene activity and silencing.

The bidirectional relationship between DNA methylation and histone modifications obscures the current understanding of the step-wise process of gene silencing. Numerous studies report contradicting results which demonstrate that either DNA methylation[198,199] or histone modifications[200,201] precede and trigger each other. Moreover, there is evidence to suggest that, in some cases, gene silencing or activation can precede changes in both DNA methylation and histone modifications[202,203]. Therefore, the potentially variable causal relationship between DNA methylation, histone modifications, and gene expression changes across different genomic regions and physiological processes remains to be further elucidated.

11.1.8 Active DNA demethylation pathway

There are also links between histone modifications and the active DNA demethylation pathway[204-206]. Though already well-established in plants, evidence for a mechanism for active DNA demethylation in mammals did not emerge until 2009, when it was first shown that TET1 catalyzes the conversion of 5mC to 5hmC[207]. Several other discoveries made in rapid succession included evidence of the same enzymatic capacity of TET2 and TET3[208], that all three TET proteins can oxidize 5hmC further to 5fC and 5caC[209], and that the protein thymine DNA glycosylase (TDG) is responsible for excision of these oxidized forms of 5mC from the genome[209]. Interestingly, the

successive oxidation of 5mC to 5hmC, 5fC, and 5caC by TET proteins stems from the fact that the structural contacts between TET proteins and the DNA substrate do not involve the methyl group itself[210]. Though the CG sequence is palindromic and typically methylated or unmethylated in both DNA strands, oxidation by TET occurs in only one strand, largely irrespective of the methylation or oxidation status of the complementary CG[211].

The next step in the active DNA demethylation pathway – following TET-mediated oxidation of 5mC – is the excision of the oxidized cytosine from the DNA by TDG. TDG is also largely insensitive to the modification status of the opposite strand[212], but is only capable of efficient excision of 5fC and 5caC and not 5hmC[213]. TDG is a glycosylase: in recognizing 5fC and 5caC, it flips the nucleotide such that it is extruded from the double-stranded DNA helix and proceeds to cleave the N-glycosidic bond that links the 9' nitrogen of the nucleobase to the 1’ carbon of the deoxyribose sugar of the backbone[214]. The excised base is retained in the binding pocket of TDG whereas the 1’ carbon of the sugar remains in the flipped position[214]. The resulting absence of a nucleobase in the DNA strand is referred to as an abasic or, more specifically, apyrimidinic site (AP site). The replacement of this AP site with an unmethylated cytosine – effectively completing the process of demethylation – requires the activity of the base excision repair (BER) pathway, which is involved in the repair of numerous lesions and is initiated by at least 11 known glycosylases with overlapping and differing lesion substrates beyond 5fC/5caC, such as 8-oxoguanine, uracil, 3-methyladenine, and others[215]. Though the BER pathway downstream of glycosylase activity is typically considered to be a common set of steps, they are mediated by different mechanisms that depend on the lesion and glycosylase type and the physiological (i.e., dividing or nondividing) state of the cell[215]. The three essential steps include cleavage of the phosphodiester bond, gap filling with the correct nucleotide by a polymerase, and ligation, ultimately resulting in a scarless correction of the double-stranded DNA[215].

The current understanding of the BER components involved in active DNA demethylation is based largely on the results of a single study which reconstituted a TET1-TDG-BER complex *in vitro* in an effort to demonstrate the complete active DNA demethylation pathway[212]. After TDG activity, the major human endonuclease, referred to as APE1 or APEX1, both displaces TDG[216] and cleaves the phosphodiester bond of the DNA backbone, which is the bond that enables the fundamental property of nucleotides to form a chain by linking the hydroxyl group at the 3' end of one nucleotide to the phosphate group at the 5' end of the following nucleotide[217]. The cleavage of the phosphodiester bond by APEX1 at the 5' end of the AP site thus generates a nick in the DNA strand, leaving 5' deoxyribose phosphate and 3' hydroxyl termini that are suitable for gap filling DNA synthesis. DNA polymerase β inserts the correct nucleotide – in this case, a cytosine, and has additional lyase activity to remove the remaining 5' deoxyribose phosphate group. In the final step of BER, DNA ligase I or III seals the DNA strand by forming a new phosphodiester bond between the existing 3' hydroxyl group and the 5' phosphate of the newly deposited nucleotide. The proteins PARP1 and XRCC1 are often also considered as BER components, not for their enzymatic activities but rather for their scaffolding function in stabilizing the interaction of both the DNA polymerase and ligase with the DNA[215]. Though this reconstituted pathway represents the current mechanistic model of active DNA demethylation, the endogenous active DNA demethylation pathway may differ in its BER components. Already, it has been demonstrated that APEX1 may not be required for active DNA demethylation and that the NEIL family of proteins can directly substitute for its activities[216]. It is important to also note that, in dividing cells, BER is not a requirement for demethylation of oxidized 5mC as 5hmC, 5fC, and 5caC are poor substrates for maintenance methylation by DNMT1 and can be rapidly lost by passive dilution[218].

The components of the active DNA demethylation are highly regulated, reflecting the dynamic nature of DNA methylation. TET and TDG mRNAs are the targets of numerous endogenous microRNAs and several post-translational modifications of TET appear to

regulate its stability and activity[219]. TET activity is also regulated at the level of genomic localization: TET proteins are enriched at active, H3K4me3-marked promoters, which may be a consequence of the binding preferences of its CXXC domain or of recruitment by its numerous protein interactors[219]. However, the physiological relevance of this active DNA demethylation pathway as a whole remains a matter of debate, despite the fact that its role across numerous cellular processes has been the subject of a large number of studies.

For example, active DNA demethylation may have a role in the dramatic global demethylation of maternal and paternal genomes shortly after fertilization, which is likely a prerequisite to both pluripotency and for the imminent gain of tissue-specific DNA methylation patterns during embryonic development[220]. While 5mC erasure in the maternal DNA occurs primarily by passive dilution in the absence of maintenance methyltransferase activity[221], the active DNA demethylation pathway appears to play some partial role in the 5mC erasure of the paternal DNA, as evidenced by a gain in 5hmC, 5fC, and 5caC in the paternal genome that is mediated by TET3[222-224]. However, the relative importance of this pathway remains unclear, as TET3 deficiency does not cause significant losses in demethylation capacity[225], though compensatory mechanisms involving TET1 or TET2 are a possible explanation. Furthermore, zygotic TDG transcript levels are negligibly low[226], suggesting that oxidized 5mC may be removed through passive dilution, without a requirement for the BER machinery. A similar wave of demethylation is observed during the differentiation of primordial germ cells[220]. Here, again, TET proteins appear to be dispensable for genome-wide demethylation[219]. However, TET1 or TET1/TET2 double knockout does result in DNA methylation changes particularly at imprinted regions (which are resist to methylation changes and typically reflect parent-of-origin methylation patterns[227]) and germline-specific genes[219], suggesting that the active DNA demethylation pathway – or, at least, oxidation by TET enzymes – is responsible for site-specific demethylation. The roles of the active DNA demethylation pathway in maintaining pluripotency are even less clear;

their combined knockout eliminates 5hmC generation but does not extensively disrupt the maintenance of pluripotency[219]. Excess methylation is indeed detectable in these knockout cells and is largely limited to enhancers and, to a small degree, promoters[219]. TDG knockout is also dispensable to the maintenance of pluripotency[209].

There is also considerable interest in the study of the active DNA demethylation pathway in the brain given the fact that adult neurons are generally considered to be nondividing[228] and thus demethylation – which has been repeatedly demonstrated to occur upon neuronal activity[229-232] – must occur by active means. Interestingly, 5hmC levels are highest in the brain compared to other adult tissues[233]. It is widely presumed that TET1 is critical for enduring neuronal-activity induced gene expression changes[63,64,232] and its ablation during conditioning exercises impairs memory formation[234]. Similarly, TET3 disruption interferes with behavioral adaptation[235] and synaptic function[236]. However, numerous confounding factors preclude the ability to ascertain whether the requirement for the functionality of TET proteins equates to a requirement for the active DNA demethylation pathway as a whole. TET proteins interact with numerous nuclear factors and affect gene expression independently of their catalytic activity[63,237] and therefore the critical gene expression changes they induce might be unrelated to their ability to participate in active DNA demethylation. Furthermore, in some studies, outcomes of genetic manipulations may reflect consequences of developmental TET deficiencies[238] rather than their roles in adult learning and memory.

Thus, a current body of knowledge that has seldom included investigations of the entire active DNA demethylation pathway and has focused more only on the consequences of the manipulation of TET protein levels raises more questions than answers. Assuming a fully functional TET-BER pathway, another question arises: are gene expression changes a result of unmethylated cytosine or do they represent BER-induced transcriptional changes[239]? More work is necessary to determine whether the presumed

pathway of active DNA demethylation by TET and BER machinery is physiologically relevant to development, differentiation, and neurological function, or whether distinct functions of TET *in vivo* inflate the importance of a functional but insignificant active DNA demethylation pathway.

11.1.9 Oxidized derivatives of the active DNA demethylation pathway and their role in gene expression

An equally formidable barrier that has obscured the relative contributions to gene expression of TET enzymes and of the potentially related but distinct replacement of 5mC with unmethylated cytosine is the fact that the historical "gold standard" methodology for the detection of demethylation is reliant on bisulfite conversion. Sodium bisulfite leads to deamination of not only unmethylated cytosine, but also 5fC and 5caC, while 5hmC is protected from deamination to a similar degree as 5mC[240]. In other words, bisulfite-sequencing does not discriminate 5mC from 5hmC nor does it differentiate 5fC and 5caC from unmethylated cytosine. Therefore, when bisulfite-sequencing is employed, oxidation of 5mC to 5hmC is masked while further oxidation to 5fC or 5caC mimics complete demethylation (i.e., replacement with unmethylated cytosine). This represents a confounding factor across numerous studies, such as many of the aforementioned investigations in the brain which claimed neuronal-activity-induced demethylation which, in using bisulfite conversion[229-231], overlooked the possibility that 5fC and 5caC are the functional epigenetic signals that lead to gene expression changes, rather than demethylation and unmethylated cytosine. Such a hypothesis would suggest a greater role of TET enzymes themselves in gene expression regulation – especially when their noncatalytic functions are also considered – than any presumed active DNA demethylation pathway.

The fact that 5hmC, 5fC, and 5caC can be readily detected in the genome[233] and are most enriched in the adult brain[233] suggests that they may indeed represent functional epigenetic signals rather than fleeting intermediates in a biochemical demethylation

reaction. This is further evidenced by studies that employed isotope labeling to demonstrate that 5hmC and 5fC are relatively stable and are not rapidly removed from the genome[241,242]. The exploration of this hypothesis can now be aided by new sequencing techniques which have been developed to probe the genomic distributions of the different cytosine modifications. Oxidative-bisulfite sequencing – the first such technique to be established – involves the chemical conversion of 5hmC to 5fC so that it is then deaminated by sodium bisulfite such that 5hmC can be effectively distinguished from 5mC[240]. The most recently published method, direct enzymatic sequencing of 5mC[243], also can distinguish 5hmC from 5mC by selective deamination of 5mC. Of course, neither technique resolves the ambiguity between unmethylated cytosine and 5fC and 5caC. For this endeavor, MAB-seq[244] utilizes the ability of the CpG methyltransferase M.SssI to methylate unmodified cytosines and its inability to methylate 5fC/5caC, which allows the discrimination of these two marks from unmodified cytosines after classical bisulfite conversion and sequencing. This method, however, suffers from a self-restricting M.SssI reaction wherein the methyl donor SAM is converted to SAH, a potent inhibitor of M.SssI: therefore, even unmodified CpGs are seldom fully methylated and 5fC/5caC rates are overestimated. MAB-seq, like oxidative bisulfite sequencing, further suffers from a requirement for a highly destructive bisulfite conversion reaction which reduces sequencing quality even in the context of abundant 5mC (discussed in Section 11.1.1); this shortcoming is aggravated by the low genomic abundance of 5fC/5caC. An alternative method, CAB-seq[245,246], utilizes selective chemical labeling of 5caC followed by bisulfite conversion, but the labeling does not reach efficiencies considered suitable for methylation detection and also relies on destructive bisulfite conversion. Instead, a bisulfite-free method called caCLEAR[247] relies on an atypical activity of engineered M.SssI (eM.SssI) to directly remove 5caC, but necessitates a set of tedious steps that include 5hmC protection by T4-BGT, methylation of unmodified cytosines by wild-type M.SssI, decarboxylation of 5caC with eM.SssI, azide-tagging, and an unconventional sequencing method that enriches for azide tags. More importantly, efficiencies of caCLEAR ranged from 49-76%. EM-seq

(introduced in Section 11.1.1) completely fails to differentiate 5mC from 5hmC, 5fC, or 5caC. Here, it is clear that no single technique used in isolation can address every possible cytosine modification state. However, the question of whether true demethylation (i.e., replacement with unmethylated cytosine) rather than oxidation is occurring in physiological examples such as learning can now be addressed by a combination of these sequencing techniques, assuming improvements in their efficiencies.

These techniques have already revealed a great deal of information about the genomic distribution of the various cytosine modifications. Though 5mC exhibits a highly tissue-specific distribution in promoters and enhancers (discussed in Section 11.1.4), the genomic distributions of 5hmC, 5fC, and 5caC appear to be even more tissue-specific. Of course, this is partially governed by the fact that 5mC is itself tissue-specific, and oxidized 5mC can only occur in the context of 5mC. The single most consistent conclusion across 5hmC mapping studies[248-255] is that 5hmC is mostly found in active gene bodies and is a stronger predictor of gene expression than gene body 5mC. A less consistently reported phenomenon across these mapping studies is that 5hmC can be found in active enhancers, as well as enhancers that are “poised” to become active. As many as one-third of 5hmC regions represent tissue-specific differentially hydroxymethylated regions[248] and 5hmC appears to accumulate with cellular age[249]. 5fC patterns[244,246,251,256-259] somewhat mimic those of 5hmC in that 5fC is found in tissue-specific active and poised enhancers and binding sites of specific transcription factors. Yet, 5fC appears to be more frequent in the CGI promoters of highly active genes, which is consistent with a positive effect on gene expression or a possible pressure for active DNA demethylation of CGIs. There appears to be little overlap between 5fC and 5caC though, interestingly, one study reported a progressive increase in histone marks indicative of active transcription as 5hmC is oxidized to 5fC and then to 5caC[246]. Like 5fC, 5caC is found in the binding sites of specific transcription factors, in unmethylated active promoters, and in active and poised enhancers[244,246,247,260,261].

Together, these findings represent a highly tissue-specific pattern of oxidized cytosine modifications and yet, they give little insight as to whether the marks function to alter gene expression or are simply indicative of sites of active DNA demethylation.

To address these two possibilities, which are not mutually exclusive, it is critical to investigate the mechanisms by which oxidized 5mC derivatives could modify gene expression. Like DNA methylation, a hydroxyl, formyl, or carboxyl group – all of which contain large, electronegative oxygen atoms – at the 5' carbon of a cytosine could dramatically affect the binding properties of transcription factors, resulting in straightforward mechanism by which oxidized 5mC derivatives could affect gene expression. Consistent with this principle, one large screen for protein interactors of 5hmC and 5fC reported numerous specific binders of 5hmC and, to a greater degree, of 5fC[262]. These included examples such as the transcription factors FOXK1 and FOXP1, DNA repair factors MHS6 and MPG, and various chromatin regulators, including all of the components of the NuRD complex, which can include MBD2 or MBD3, HDACs and other proteins. An independent study also identified the MBD3/NuRD complex as a binder of 5hmC that regulates expression of 5hmC-containing DNA[263]. Interestingly, the results of this thesis (Chapter 3) also support a role for MBD3/NuRD in the recognition and regulation of oxidized cytosines and reveal the recruitment of TDG by this complex. Other data has suggested that 5caC increases the binding affinity of CTCF to genomic sites that otherwise show suboptimal binding[261]. It is also likely that 5hmC, 5fC, and 5caC could inhibit certain interactions. In light of this constantly evolving body of evidence, considerably more work is required to improve detection methodologies, clarify the mechanistic relationship between oxidized 5mC derivatives and gene expression, and ascertain the relative contributions of these marks compared to active DNA demethylation across a wide range of physiological processes.

11.2 The causality of DNA methylation in gene expression changes

The identification of DNA methylation changes and gene expression changes which may or may not associate with different diseases and physiological conditions – such as cancers and memory formation – in itself reveals no causal relationship between the change in methylation and gene expression or the phenotypic outcome. The observation that, for example, a tumor suppressor is methylated and transcriptionally silenced in cancer, is not evidence of the fact that the DNA methylation change is driving the transcriptional repression. This is reflected in the fundamental scientific principles of causation and correlation. It may well be that, across different scenarios, the opposite is true and transcriptional changes direct DNA methylation changes[202,203] or, in other circumstances, DNA methylation changes are non-functional bystanders of a conserved and cross-talking silencing machinery wherein DNA methylation is only sometimes functional (e.g., at specific developmental times or genomic locations) while other covarying factors (i.e., repressive histone marks) represent the functional silencing modification[264]. Understanding the causal relationship between the DNA methylation state and gene expression at specific genomic loci and in specific physiological contexts is critical to understand the mechanisms that regulate gene expression and contribute to disease and to thereafter assess whether specific DNA methylation changes should constitute therapeutic targets.

11.2.1 Global modifiers of DNA methylation

Many studies which report an association between DNA methylation and gene expression changes do indeed attempt to establish causality. In this context, causation can be demonstrated by the manipulation of the methylation status of a site of interest – for example, a specific promoter – and subsequent quantification of any resulting changes in gene expression, in the absence of confounding variables. A commonly utilized approach to demonstrate causality in DNA methylation research is through pharmacological or genetic manipulation of endogenous DNA methylation levels. Pharmacological inhibition of DNA methylation has been classically achieved by two agents: SAM and 5-aza-2'-deoxycytidine. Treatment of cells with the methyl donor SAM

has been used to successfully increase methylation of a promoter of interest and often can lead to reduced expression of that gene, apparently serving as a demonstration of causality[265,266]. However, SAM is also the methyl donor for every methylation reaction in the cell and thus affects the function of over 200 methyltransferases[267] whose targets span from DNA to RNA, histones, and other proteins and SAM further participates in numerous metabolic synthesis pathways such as polyamine and cysteine synthesis[268]. Thus, treatment with SAM does not isolate DNA methylation of a specific promoter as an independent variable and invalidates the fundamental requirements for establishing causality.

5-aza-2'-deoxycytine, on the other hand, has the opposite effect, decreasing DNA methylation. 5-aza-2'-deoxycytidine is a synthetic cytidine analog in which the 5' carbon of cytidine is replaced by a nitrogen atom and it, upon incorporation into the DNA, exploits the catalytic mechanism of DNMTs to inhibit their function[269]. The catalytic mechanism of DNMTs involves a crucial step wherein a thiol group of a cysteine in the active site of the enzyme performs a nucleophilic attack on the carbon-6 (C6) of the target cytosine, forming a transient covalent bond[269]. This results in activation of the C5 atom for electrophilic attack by SAM and thus the methyl group is added to the cytidine. This methyl group addition is then succeeded by the elimination of the C5 proton and the resolution of the covalent intermediate[269]. When 5-aza-2'-deoxycytidine is instead in the active site of a DNMT, the transient covalent intermediate between C6 and the enzyme is still formed[270] and the methyl group is transferred to the nitrogen[271]; however, with no additional hydrogen atom present on the nitrogen, the transient bond cannot be resolved and the DNMT remains covalently bound to 5-aza-2'-deoxycitidine[269]. At low 5-aza-2'-deoxycytidine concentrations, this irreversible complex traps DNMTs and thus diminishes the pool of available DNMTs, reducing their activity elsewhere by as much as 95%[272] and leading to rapid and considerable passive loss of DNA methylation in dividing cells[272,273], typically including that of any specific promoter under study and often accompanied by an increase in expression from that promoter. To this end, 5-aza-

2’-deoxycytidine treatment is an exceedingly common method to demonstrate that the expression of a specific gene is regulated by promoter DNA methylation[274-279]. At high 5-aza-2’-deoxycitidine concentrations, the large amount of DNMT-DNA adducts where 5-aza-2’-deoxycytidine was incorporated in the DNA inhibits polymerase movement along the DNA strand and thus interferes with DNA synthesis, leading to growth arrest and cell death[280]. Again, an issue emerges in isolating DNA methylation as a variable: by affecting DNMTs and not DNA methylation, all of the aforementioned interacting partners of DNMTs, such as modifiers of histone marks, are also affected by 5-aza-2’-deoxycitidine and thus any observed transcriptional effects of demethylation can and do originate from sources other than promoter demethylation[281,282]. Moreover, 5-aza-2’-deoxycitidine-DNMT adducts trigger a DNA damage response that is known to affect local gene expression by both the adduct itself and its repair and results in numerous changes in histone modifications[239].

The latter issue of DNA damage may be resolved by a new generation of non-covalent DNMT inhibitors which are not nucleoside analogs[283], but this practice has not yet been widely adopted in causality research. Instead, a commonly utilized and potentially less confounded approach to DNMT inhibition involves genetic techniques. The reduction of DNMT protein levels can be achieved by genetic knockout (e.g., using CRISPR/Cas9, which is described in Section 11.2.3) or by degradation of DNMT mRNA by the expression of small interfering RNAs (siRNAs) targeting the DNMT sequence, which interact with the endogenous miRNA processing protein complex RISC and, together, cleave the DNMT mRNA or inhibit its translation[284]. Genetic disruption of DNMT1 in this manner can lead to global DNA demethylation, whereas targeting of DNMT3A or DNMT3B may be relevant for specific genes that they regulate[285]. This approach still has the same caveats as 5-aza-2’-deoxycytidine in that it fundamentally represents DNMT inhibition rather than direct DNA demethylation and is thus susceptible to be confounded by effects mediated by DNMT interactors. There is also a further issue of off-target inhibition: siRNA-mediated silencing is tolerant of a few mismatches in the

siRNA, and thus expression of genes other than the DNMTs may be interfered with, confounding the relationship further[284]. Still, this approach has been widely used to demonstrate the causal effect of DNA methylation at specific promoters on gene expression[286-289].

A uniting pitfall of all of these global modifiers of DNA methylation is that they are not site-specific. Hypo- or hypermethylation of the entire genome accompanies that of any gene of interest and represents a major additional confounding factor. It is now widely established that 5-aza-2'-deoxycytidine[290-292] and genetic DNMT1 depletion[291] can lead to activation of gene expression from both unmethylated and methylated promoters and without any changes in the DNA methylation levels of that promoter. This can occur by effects of 5-aza-2'-deoxycytidine on histone modifications and by a hierarchical activation of expression of other transcription factors and proteins that then regulate expression from this promoter[281,291]. Therefore, the narrow analysis of a gene of interest in the presence of any global modifier of DNA methylation – in addition to other confounds specific to each technique – is inherently not sufficient to establish a causal relationship between the methylation and expression of a particular gene.

11.2.2 Promoter-reporter analyses

Another approach for assessing causality in DNA methylation research is through the use of methylated promoter-reporter constructs. This is achieved by a series of experimental steps. First, the promoter of a gene of interest is cloned into a reporter plasmid such that the promoter is upstream of and drives expression of a reporter gene, such as luciferase or green fluorescent protein (GFP). This construct is then methylated *in vitro* by an efficient bacterial methyltransferase, which is typically M.SssI due to its lack of sequence specificity beyond the CG dinucleotide or, less frequently, M.HhaI, which methylates CG cytosines in a GCGC context. The unmethylated and methylated forms of the construct are then transfected into cells and expression of the reporter can be detected and quantified as a function of methylation. This common approach has

been widely used throughout the field since its use in the earliest demonstrations of causality by Kruczek and Doerfler[120] and typically results in the expected decrease in reporter expression upon methylation of any promoter which contains CG sites near regulatory elements[120,133,279,293-299].

Methylated promoter-reporter assays allow the manipulation of the DNA methylation level of a single promoter of interest and therefore reduce the confounds involved with global DNA methylation modifiers at two levels: the independent variable is now DNA methylation rather than DNMT inhibition and the change in methylation occurs at a single promoter, rather than genome-wide. Moreover, promoter-reporter assays also allow a further capacity to precisely interrogate the causal role of methylation of specific CG sites or stretches of CGs rather than that of the entire promoter itself. This can reveal precise transcription factor interactions that are disrupted by DNA methylation: for example, one study demonstrated that while methylation suppresses 90% of expression from the insulin promoter, methylation of only a single CG within a cAMP responsive element (CRE) in the insulin promoter is responsible for 50% of the reduction in gene expression[133]. Such distinctions can be achieved by either cloning different fragments of the same promoter into reporter constructs and conducting parallel methylated promoter-reporter assays[133] or by a method known as “patch methylation”, wherein different fragments of the promoter produced by restriction enzyme digest are methylated or not in different combinations and are then ligated into the reporter construct to reconstitute the promoter[300,301], or by related techniques involving the cloning of methylated synthetic oligonucleotides[302].

Yet, methylated-promoter reporter assays are not completely unconfounded. The DNA analyzed in these assays is not endogenous; it is an artificial construct that copies the sequence of an endogenous promoter but is located in a circular plasmid, has arbitrary termini at which it was cloned, and is transfected into cells. This raises several questions concerning the degree to which the methylation and expression relationships

observed with this approach reflect true endogenous relationships between expression and methylation of genomic DNA. It is possible that the same promoter would not exhibit the same changes in expression if the methylation had occurred in the endogenous context, which includes additional regulatory sequences at either side, simultaneous interactions with transcription factors, larger chromatin structural factors, and a landscape of coexisting regulatory histone modifications. Furthermore, it remains unclear whether expression of ectopic DNA is fundamentally comparable to endogenous gene expression or whether it instead represents some wholly unrelated response to invading foreign DNA. Moreover, difficulties in patch methylation of more than two promoter fragments and questions about the scientific validity of investigating contributions to gene expression of independent promoter fragments in the absence of surrounding elements complicate site-specific analyses of DNA methylation with promoter-reporter assays. Newer tools capable of site-specific and CG-specific editing of DNA methylation in the endogenous context could resolve these shortcomings and allow a better understanding of causality.

11.2.3 An overview of targeted DNA methylation editing techniques

The most recently developed approach to study the causal role of DNA methylation in gene expression involves the site-specific manipulation of DNA methylation levels in the genome of living cells. While “gene editing” refers to the targeted manipulation of the DNA sequence, the same tools have been repurposed for “epigenetic editing” or, more specific to this thesis, DNA methylation editing. A tool for site-specific epigenetic editing typically must consist of two components: an enzymatic component with epigenetic-modifying activity and a targeting component – a domain that can bind a specific DNA sequence so that the epigenetic activity can be targeted to specific genes or genomic locations.

At present, targeting can be achieved by one of three categories of targeting domains: zinc-fingers (ZFs), transcription activator-like effectors (TALEs), or CRISPR/dCas9[303].

ZF targeting domains are composed of several naturally occurring zinc finger motifs – found in natural DNA-binding proteins – which are known to recognize and bind to a specific predefined triplet of nucleotides. By combining multiple zinc finger motifs together (i.e., by cloning), a custom DNA-binding domain can be created with high specificity for a desired target sequence[303]. TALEs have a similarly modular structure consisting of repeat domains though, unlike ZFs, each repeat recognizes only one specific nucleotide. TALEs also have a natural origin in certain species of pathogenic bacteria (mostly Xanthomonads) in which they work to bind plant DNA to activate expression of plant genes that aid in bacterial infection. By arranging the appropriate repeats in a specific order, a synthetic TALE array can be designed to recognize a particular DNA sequence[303]. Unlike ZFs and TALEs, CRISPR/Cas9 is an RNA-guided targeting system that does not require protein engineering. CRISPR/Cas9 was co-opted from bacterial adaptive immune systems and involves a ~20-bp targeting sequence referred to as a guide RNA (gRNA) that associates with the Cas9 protein. Together, this complex scans the genome for sites that are complementary to the gRNA sequence and, upon recognition of a target site, binds to and cleaves both DNA strands at the target site[303]. The two nuclease domains of Cas9 can each be inactivated by single point substitutions to create a nuclease-dead Cas9 (dCas9) that can serve as a site-specific targeting system analogous to ZFs and TALEs[304].

In the case of TALEs, the highly homologous repeats of 33-35 amino acids are prone to recombination and thus encounter difficulties at every relevant stage: synthesis, cloning, and introduction into target cells[303]. They do, however, possess a key advantage over ZFs, in that they can be modified to target nearly any sequence. Conversely, there is no available library of 64 zinc-fingers that are capable of binding each possible nucleotide triplet[303]. In the case of both tools, the new proteins require extensive validation after construction in order to evaluate its on- and off-target effects. All of these deficiencies render the tools time-consuming and expensive to use. For research purposes, CRISPR/dCas9 is by far the simplest to re-target, requiring only a change of the gRNA

sequence rather than more complicated *de novo* protein design necessary for new TALEs or ZFs. CRISPR/Cas9-based systems thus also have a major advantage of being able to used for genome-wide screens by using a library of gRNAs that target every human gene[305,306]. Thus, TALEs and ZFs have largely been abandoned in the DNA methylation editing field and the epigenetic editors discussed herein will be limited to those that utilize CRISPR/dCas9 as the targeting component. It is important to add that, while this ease of use has resulted in the widespread adoption and preference for CRISPR/dCas9-based architecture in the research laboratory setting, TALEs and ZFs remain highly useful in the clinic, where the time-consuming development of a custom protein that targets a single locus is acceptable when the therapeutic goal is the manipulation of a single known disease-causing gene or locus[307,308].

11.2.4 Enzymatic epigenetic engineering for targeted DNA methylation and its shortcomings

Currently, the enzymatic component of any epigenetic editor for targeted DNA methylation must be a DNA methyltransferase – an enzyme capable of catalytic addition of methyl groups to DNA. Targeted DNA methylation in the mammalian genome using CRISPR/dCas9 was first demonstrated in 2016 by four independent groups: all four approaches were identical in their reliance on the fusion of dCas9 to the de novo DNA methyltransferase DNMT3A (or its catalytic domain)[309-312]. Though these studies provided an excellent proof-of-principle, the major downside was that efficient methylation required prolonged high expression of dCas9-DNMT3A and declined rapidly after termination of dCas9-DNMT3A expression. In 2017, two studies reported increased efficiencies of targeted methylation either by the addition of the DNMT3A-stimulating protein DNMT3L as part of the methylation effector domain[313] or by replacing DNMT3A entirely with the potent bacterial methyltransferase M.SssI[314].

While complete and persistent targeted methylation continues to be elusive, the considerably more serious drawback underlying all targeted methylation strategies is

the pervasive off-target activity of the methyltransferase effector domains, which are overexpressed in all such strategies and are capable of nonspecific genome-wide methylation independent of the targeted dCas9 that they are fused to. The observation that methyltransferase constructs possess a nonspecific activity independent of targeting that renders them inadequate for precise epigenetic editing has been widely accepted since before the development of CRISPR/dCas9 technology[315] and has continued to be consistently demonstrated to be the case with CRISPR/dCas9. Even in the least efficient targeted methylation strategies – fusions of the catalytic domain of DNMT3A to dCas9 (dCas9-DNMT3A-CD) – the expression of dCas9-DNMT3A-CD causes global methylation changes in cells that are highly similar to those observed upon the expression of DNMT3A alone[316]. dCas9-M.SssI exhibits a similar and potent nonspecific activity[314].

There have been numerous efforts to improve specificity. Two independent groups reported similar strategies which involved activity-reducing point mutations in the DNMT3A (R887E)[317] or M.SssI (Q147L)[314] components of dCas9-based epigenetic editors, both claiming increased methylation specificity. For DNMT3A (R887E), off-target methylation, though reduced, was still detectable: targeted analysis of an arbitrary off-target region in the VEGFA promoter consistently revealed gRNA-independent off-targeted methylation of the R887E mutant, suggesting that the accompanying genome-wide methylation (MBD-seq) analyses (which also reported mild off-target methylation) might be underpowered and underestimate the off-target methylation events that might be detected with a more powerful genome-wide DNA methylation analysis method. For M.SssI (Q147L), the authors analyzed candidate gRNA off-target sites of the gRNA and reported no off-target methylation – yet, this off-target analysis strategy is insufficient in that it does not assess the aforementioned off-target methylation independent of dCas9 by the methyltransferase domain as has been extensively demonstrated. Though the authors also performed reduced-representation bisulfite sequencing, they correctly conclude that a failure to detect off-target effects with these approaches does not

equate to an absence of off-target effects and, indeed, a more recent study presented evidence that the Q147L mutation in M.SssI does not reduce its nonspecific activity[318], instead only reducing its catalytic activity, which suggests that in the original study, the Q147L mutation leads to off-target methylation that is below the threshold of detection. Off-target detection is highly dependent on the power and comprehensiveness of the detection method: it is likely that deep whole-genome methylation sequencing – which was not used in either study – would reveal the true extent of off-target methylation of these engineered epigenetic editor variants.

Another approach reported reduction of the nonspecific methylation by recruitment of multiple DNMT3A domains by the SunTag system – which employs a single-chain antibody fusion to DNMT3A while multiple copies of the peptide repeats recognized by this antibody (epitopes) are fused to dCas9 – rather than direct fusion to dCas9[319,320]. Still, this strategy continues to require overexpression of the methyltransferase domain and nonspecific methylation was still reported. Furthermore, contradictory results from an independent group reported no observed increase in specificity of the SunTag approach[317].

A final strategy to reduce off-target methylation of epigenetic editors is to split the methyltransferase domain into two components such that the complete methyltransferase domain is reconstituted only at the targeted site (either by targeting the two split domains separately or by targeting one of the split domains to the target site and constitutively expressing, without any targeting, the other split domain). Though there is some evidence that this split methyltransferase approach improves non-specificity compared to dCas9-DNMT3A fusion proteins, the data is limited to only a few targeted sites and lacks whole-genome methylation analysis[321]. Moreover, there are contradictory reports that split methyltransferase strategies exhibit similar efficiencies in methylation of target and non-target sites[315]. Off-target methylation activity thus appears

to be an inevitable correlate of targeted methylation as long as epigenetic editors remain invariably dependent on the overexpression of methyltransferase domains.

An additional point to keep in mind regarding dCas9-methyltransferase fusions is that human methyltransferases have evolved to interact with dozens of nuclear proteins to regulate transcription, which could lead to unintended and confounding consequences of the epigenetic editor both at the target site and throughout the genome. Similar to previous discussions of 5-aza-2'deoxycytidine and DNMT inhibitors, the independent variable in these methods is DNMT localization and is not direct DNA methylation modulation. DNMT3A, for example, interacts with EZH2[196] and p53[322] and many other proteins and therefore can modify expression independently of any DNA methylation changes. These interactions are likely to persist even when only the catalytic domain of DNMT3A is used but are likely less relevant for epigenetic editors relying on the non-human methyltransferase M.SssI.

Exceptionally, a new approach to induce methylation in an unmethylated promoter without using enzymes involves the disruption of an unmethylated CG island by integration of a fragment of CG-less DNA into embryonic stem cells using CRISPR-mediated targeting and recombination[323]. This method triggers methylation of the disrupted CG island and, after inducing methylation, the CG-less fragment is removed by Cre-Lox or Piggybac transposase mediated recombination. The introduced methylation can be stable and trans-generationally heritable in mice, with resulting changes in gene expression and phenotype[324]. While this method holds some value for research purposes, it requires a high cellular de novo methylation activity that restricts its use mostly to embryonic stem cells[325]. Furthermore, this approach can only induce regional methylation, not site-specific methylation. Finally, in its current form, the technique involves several genetic changes inherent to the recombination (i.e., numerous variants between two different alleles are turned into nonvariable sites bearing the genotype of one of the two alleles) as well as those potentially from the use

of catalytically active CRISPR/Cas9, which may also introduce off-target mutations[326] and genomic alterations[327] that could confound interpretation of results and limit its utility.

11.2.5 Enzymatic epigenetic engineering for targeted DNA demethylation and its shortcomings

The characteristics of targeted enzymatic DNA demethylation techniques parallel those of techniques for targeted methylation. In humans, active DNA demethylation is initiated by the TET family of enzymes, which catalyze the conversion of methylated cytosines to a series of more oxidized forms that are eventually excised from the genome and replaced by unmodified/unmethylated cytosines by DNA repair machinery[219]. Therefore, though TET proteins are not biochemically demethylases, they are classically used as the enzymatic component of CRISPR/dCas9-based targeted demethylation techniques (**Figure 1**).

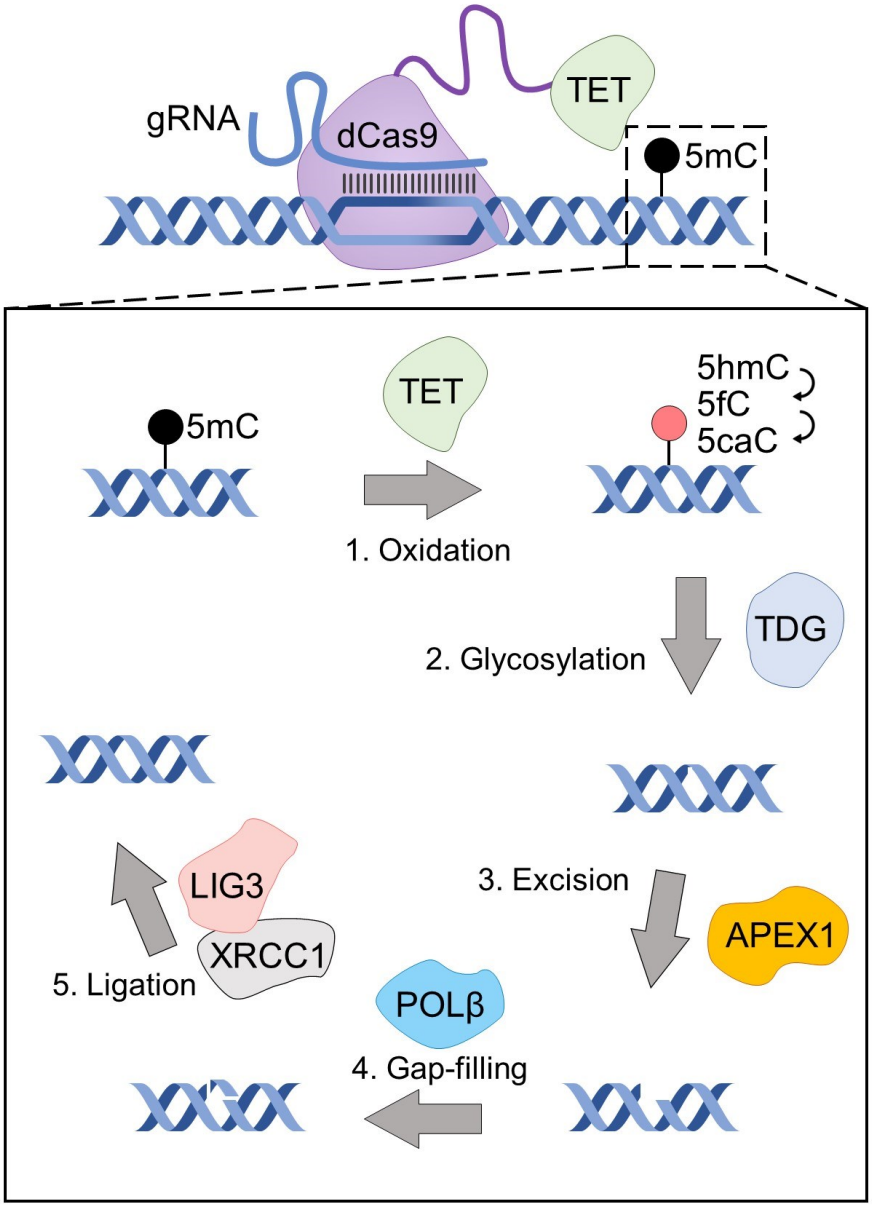


Figure 1. DNA demethylation by dCas9-TET. A schematic diagram of dCas9-TET-based targeted DNA demethylation, highlighting the numerous steps and enzymes required for a methylated CG to be converted to an unmethylated CG by this epigenetic editing tool.

In 2016, four independent groups developed CRISPR/dCas9-based systems for targeted DNA demethylation: those that fuse[311,328] or recruit[329,330] the catalytic domain of TET1 with CRISPR/dCas9. Interestingly, despite the ability of all three TET family proteins (TET1, TET2, TET3) to oxidize methylated CGs[69], fusions of dCas9 to the catalytic domain of TET2 and TET3 have only been presented in one thesis[331] and one article[332], respectively. The general absence of other known efficient demethylases from humans or other organisms has resulted in comparatively reduced innovation in the demethylation editing field (compared to methylation editing, described above) and has yielded a much simpler landscape of epigenetic editing tools for targeted demethylation than those for methylation. However, there is one strategy for targeted demethylation that does not invoke TET enzymes: ROS1, a glycosylase from *Arabidopsis*, is able to directly initiate replacement of methylated CGs without need for an initial oxidation step. To this end, dCas9-ROS1[333] has been successfully used to demethylate CGs by way of direct glycosylation.

Unlike the tools for targeted DNA methylation, there have been no systematic comparisons of CRISPR/dCas9-based demethylation tools. As discussed previously, TET proteins have numerous non-catalytic activities and gene expression activation by a catalytic mutant of TET has been previously reported[63]. It was also previously reported that the transcriptional effects of TET depletion in cells without all three DNMTs are similar to those in wild-type cells, suggesting significant methylation-independent activities of TET[334]. Similarly, TET1 was shown to regulate H3K27 modification independent of its catalytic activity as a catalytic TET1 mutant expressed in embryonic stem cells or mice restored the normal pattern of H3K27me3 and normal differentiation which were deregulated in a TET knockout model[237]. Therefore, DNA demethylation by dCas9-TET is not anticipated to be the only effect and these methods are therefore not suited to addressing the causality of DNA demethylation in gene expression. Off target effects of dCas9-TET were also reported[335]. Epigenetic editing with a tethered TET enzyme is therefore confounded by (1) the methylation-independent activities of TET as

seen by the efficient gene activation with a catalytic TET mutant and its established DNA methylation-independent transcriptional and histone modification effects, (2) like dCas9-DNMT3A, the potential genome-wide untargeted effects of the tethered TET enzyme, and (3) the fact that TET proteins are dioxygenases and not demethylases and therefore it is difficult to discriminate whether epigenetic effects are caused by newly demethylated or newly oxidized CGs, which impact transcription through numerous gains or losses of interactions[336]. Moreover, TET mediated demethylation requires base excision repair by TDG and other proteins which might compromise the integrity of the DNA and possibly indirectly affect gene expression. Therefore, TET-based epigenetic editing tools – in contrast to DNMT-based epigenetic tools, which are methylating enzymes – are further confounded by the fact that they make use of an enzymatic activity that is not directly demethylating the DNA.

11.3 Research goals and scope of the thesis

The previous sections have emphasized certain insufficiencies in the field of DNA methylation research. The first major shortcoming is the lack of causal relationship to gene expression attributed to specific instances of DNA methylation and therefore a poorly defined understanding of the functional consequences of DNA methylation in general. This is largely due to the complete absence of a simple, efficient, and unconfounded DNA methylation editing tool which would be able to modify the methylation status of specific CGs and allow researchers to assess the consequences. Moreover, a precise and epigenetic editor has potential applications in the clinic. Disease-causing disruptions to gene function could occur due to changes in the primary nucleotide sequence, which could be corrected through genetic editing, as well as seemingly by epigenetic dysregulation, which could potentially be corrected through epigenetic editing. The fundamental difference between epigenetic and genetic editing is that epigenetic programming is reversible, and “correction” of an epigenetic program does not require alterations of the integrity of the DNA sequence and is thus more appealing as a therapeutic strategy than gene editing.

Though epigenetic editing can target either histone or DNA modifications, DNA modifications are base-specific and can provide higher sequence-specific resolution than histone modifications, which can cover a wider region. There is a significant body of evidence that demonstrates DNA methylation alterations in various diseases, such as cancer[337-340], autoimmune disease[341,342], cardiovascular disease[343-346], and metabolic disease[342,347]. If it is indeed demonstrated that specific alterations are causal, targeted epigenetic editing could potentially reverse the disease-associated DNA methylation profiles: in other words, targeted epigenetic editing is well-suited to address both causality and correction. Targeted epigenetic editing may also have applications in cell therapy and immune therapy, as in, for instance, trans-differentiation of liver cells to produce insulin through demethylation of the insulin gene and pancreatic-specific transcription factors[133,346] or silencing of checkpoint inhibitors in patient T cells[348,349] to enhance immunotherapy. However, there are still numerous barriers to overcome.

The first obstacle is the continued absence of this simple, efficient, and unconfounded DNA methylation editing tool. A major aim of this thesis was to develop such a tool and one that could exploit the benefits of the newly discovered CRISPR/Cas9 system. At the time when the work in this thesis was started, there were no published CRISPR/Cas9 based epigenetic editing tools for DNA methylation or for DNA demethylation. However, a tool for targeted DNA methylation was never considered, as tools for methylation are invariably reliant on DNMTs and thus are invariably confounded. As, in theory, dCas9-TET as an DNA demethylating tool is also confounded at multiple levels, it cannot fully address the causal role of DNA methylation or be used as a specific demethylating agent. In addition to these limitations, any epigenetic editing tool that consists of an enzyme tethered to dCas9 would modify a region of DNA of varying sizes based on processivity, the steric parameters, chromatin structure, and the flexibility of the tethered enzyme and it therefore cannot be targeted to a specific CG. Here, we supposed that DNA demethylation could instead be triggered at targeted sites by a simple steric

interference with DNMT1 by bound dCas9 alone, such that these sites are inaccessible to DNMT1 and methylation only at these sites is diluted with every round of cell division and DNA replication, effectively producing DNA demethylation (**Figure 2**).

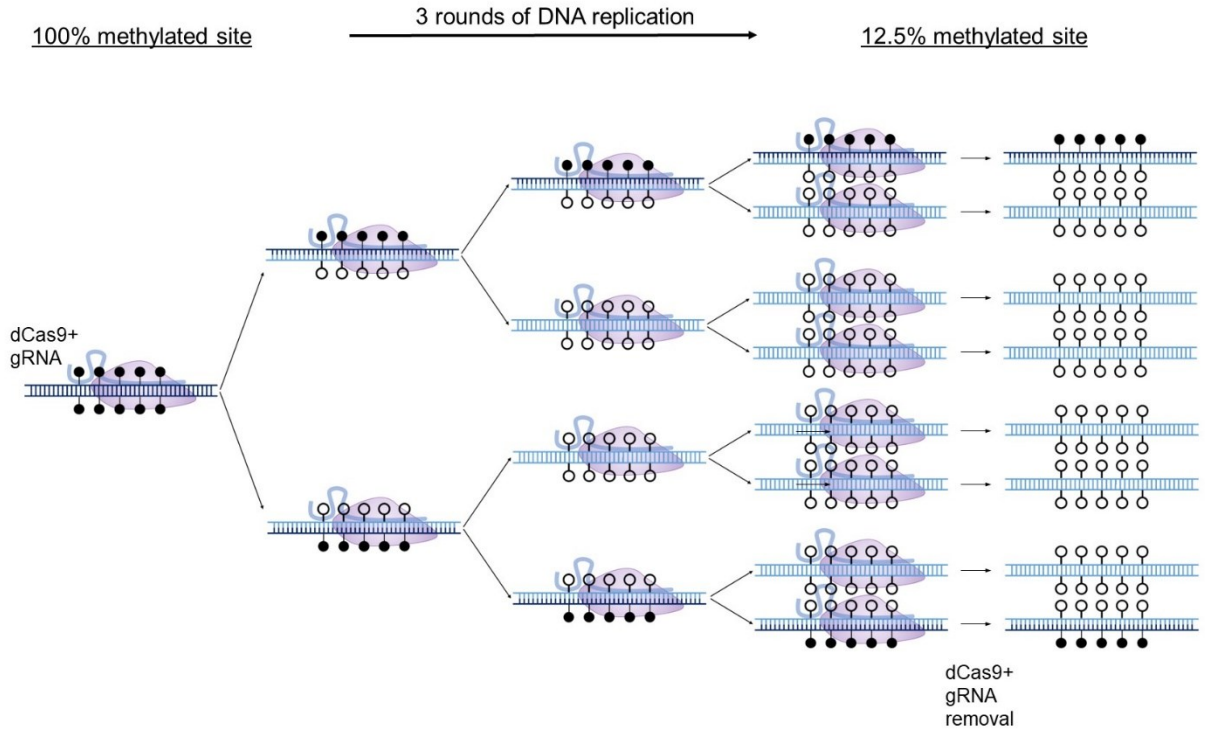


Figure 2. DNA demethylation by dCas9 steric hindrance. A schematic diagram depicting the targeted DNA demethylation induced by the steric hindrance method in dividing cells. Parental DNA strands, shown in dark blue, are diluted as the cells divide. In the presence of a fully effective targeted interference of dCas9 with DNMT1, after 3 rounds of cell division wherein methylation levels halve with every round due to a lack of methylation of nascent DNA strands (light blue), a target site which was originally 100% methylated would be effectively demethylated to 12.5%, though, in practice, more rounds of cell division should be included to further reduce methylation. dCas9 and

gRNA expression can then be terminated, leaving the unmethylated target site exposed to potential interacting proteins in the nucleus.

Therefore, more specifically, the aims of Chapter 2 of the thesis:

1) Demonstrate that dCas9-TET tools, which had become widely accepted in the field for demonstrations of causality, are not suited to assess causality.
2) Develop instead an enzyme-free CRISPR/dCas9-based epigenetic tool, characterize its activity, optimize its efficiency, and assess its specificity so that it may be adopted as a standard protocol for assessing causality in the field.
3) Demonstrate a proof-of-concept application of this new tool at several methylated promoters across human and mouse cell lines in order to assess the causal relationship between DNA demethylation and gene expression.
4) Demonstrate the utility of this method in a clinically relevant scenario.

The second major shortcoming addressed in this thesis is the obscurity of the TET-catalyzed active DNA demethylation pathway, including its protein components and its oxidized 5mC intermediates. Numerous issues raised in previous sections include inefficient and destructive oxidized 5mC detection technologies, a resulting unclear relationship between oxidized 5mC modifications and gene expression, unclear physiological contributions of BER machinery and the entire endogenous active DNA demethylation, and whether gene activation by TET precedes and is necessary demethylation, or vice versa. Therefore, the aims of Chapter 3 of this thesis were to:

1) Develop a novel, bisulfite-free technique for the detection of oxidized cytosines that is simple to use and efficient.
2) Use this new technique to perform a more robust study of the dynamics of the active DNA demethylation pathway using transfected oxidized promoter-reporter DNA in tandem with a series of genetic knockouts or mutations of its protein

components and genetic manipulations of the promoter-reporter itself to study the relationships between oxidized 5mC derivatives, active DNA demethylation, and gene expression.

3) To understand whether MBDs – particularly the previously implicated MBD3/NuRD complex – could participate in the active DNA demethylation pathway or oxidized 5mC signaling.
4) To use this new technique and any discoveries from the previous aims to study active DNA demethylation and oxidation in an *in vivo* context – in the mouse cortex, which has high levels of these modifications.

Chapter 2: Unraveling the functional role of DNA demethylation at specific promoters by targeted steric blockage of DNA methyltransferase with CRISPR/dCas9

Sapozhnikov, D.M., Szyf, M. Unraveling the functional role of DNA demethylation at specific promoters by targeted steric blockage of DNA methyltransferase with CRISPR/dCas9. *Nat Commun* 12, 5711 (2021). https://doi.org/10.1038/s41467-021-25991-9

12.1 Abstract

Despite four decades of research to support the association between DNA methylation and gene expression, the causality of this relationship remains unresolved. Here, we reaffirm that experimental confounds preclude resolution of this question with existing strategies, including recently developed CRISPR/dCas9 and TET-based epigenetic editors. Instead, we demonstrate a highly effective method using only nuclease-dead Cas9 and guide RNA to physically block DNA methylation at specific targets in the absence of a confounding flexibly-tethered enzyme, thereby enabling the examination of the role of DNA demethylation per se in living cells, with no evidence of off-target activity. Using this method, we probe a small number of inducible promoters and find the effect of DNA demethylation to be small, while demethylation of CpG-rich FMR1 produces larger changes in gene expression. This method could be used to reveal the extent and nature of the contribution of DNA methylation to gene regulation.

12.2 Introduction

DNA methylation is broadly involved in transcriptional regulation across a vast number of physiological and pathological conditions[350]. For nearly half a century, it has been widely documented that the presence of methyl groups on the fifth carbon of cytosines in the context of CpG dinucleotides within promoters is associated with transcriptional repression[351]. This is considered to be a crucial epigenetic mark as deviations from the tightly-regulated and tissue-specific developmental patterns have been implicated in conditions as diverse as cancers[352], suicidal behavior[353], and autoimmune diseases[341]. Yet, these studies also exemplify a fundamental challenge in the field: the persistent inability to attribute causality to a particular instance of aberrant DNA methylation. The issue of whether DNA demethylation is the driver of relevant transcriptional changes continues to be a source of controversy and is magnified by multiple studies suggesting that changes in gene expression and transcription factor binding can in some cases precede DNA demethylation[202,203,354-357]. The answers to this set of questions would reveal whether a particular DNA methylation state is only a marker for a particular condition or whether it is plays a critical role in the pathophysiological mechanism.

In the case of DNA methylation, unconfounded manipulation of the methylation state of a CpG or region of CpGs in isolation remains a challenge: genetic (DNA methyltransferase knockdown) and pharmacological (5-aza-2'-deoxycytidine and S-Adenosyl methionine) hypo- or hyper-methylating agents cause genome-wide changes in methylation[40,280,285,358,359], confounding conclusions by countless concurrent changes throughout the genome in addition to any region under study. A more specific approach to assessing causality involves comparing the abilities of *in vitro* methylated and unmethylated regulatory sequences to drive reporter gene expression in transient transfection assays. However, this is an artificial system and a simplification of the complex chromatin architecture at the endogenous locus, and therefore the effects of methylation in the context of an artificial promoter-reporter plasmid may misrepresent those that would occur under physiological conditions.

More recently, the TET dioxygenases – which oxidize the methyl moiety in cytosine and can lead to passive loss of methylation by either inhibiting methylation during replication

or through repair of the oxidized methylcytosine and its replacement with an unmethylated cytosine – were targeted to specific sites using a fusion of TET dioxygenase domains to catalytically inactive CRISPR/Cas9 (dCas9) [209,328-330]. However, this method still introduces several confounding factors that preclude causational inferences, such as the fact that oxidized methylcytosines are new epigenetic modifications that are not unmethylated cytosines[360-366] and the fact that TET has methylation-independent transcriptional activation activity [367,368].

We propose and optimize instead an enzyme-free CRISPR/dCas9-based system for targeted methylation editing which we show is able to achieve selective methylation *in vitro* and passive demethylation in cells through steric interference with DNA methyltransferase activity. We map the size of the region of interference, optimize the system for nearly complete demethylation of targeted CpGs without detectable off-target effects, and analyze the transcriptional consequences of demethylation of genetically dissimilar regions across several human and mouse genes. In doing so, we provide evidence that DNA demethylation at proximal promoters increases gene expression in some instances but not others, that it does so to varying degrees depending on the genomic context, and that demethylation may facilitate responses to other factors. Most importantly, we report a simple tool for investigations into the effects of DNA demethylation that can be applied with ease and in multiplexed formats to examine the vast existing and forthcoming correlational literature in order to distinguish causational instances of DNA methylation and begin to develop a fundamental understanding of this biological phenomenon on a foundation of causality.

12.3 Results

CRISPR/TET-based approaches confound the causal relationship of DNA methylation and transcription

To develop a tool for site-specific DNA methylation editing, we elected to study the murine interleukin-33 (*Il33*) gene. The distance between individual CpGs and sets of

CpGs within its canonical CpG-poor promoter provides a simple starting point that enables specific CpG targeting in order to evaluate the impact of discrete methylation events on gene transcription (**Figure 1A**). The promoter is highly methylated in NIH-3T3 cells (**Supplementary Figure 1A**) and upon treatment of cells with the demethylating agent 5-aza-2'-deoxycytidine, CpGs adjacent to the transcription start site (TSS) are demethylated (**Figure 1B**) and gene expression is moderately induced (**Figure 1C**). However, this classical response to 5-aza-2'-deoxycytidine also emphasizes the shortcomings of this common approach in DNA methylation research: (1) multiple CpGs in the promoter are demethylated, so it remains unclear which sites of methylation contribute to transcriptional inhibition, and (2) the global genomic consequences of 5-aza-2'-deoxycytidine treatment result in the induction of expression of several putative and experimentally validated *Il33* transcription factors (**Figure 1D**), exemplifying the possibility that demethylation of the *Il33* promoter may not be the event responsible for upregulation of the gene. This demonstrates a need for an accurate and specific targeted methylation editing technology that can properly interrogate the fundamental question of the causal relationship between DNA methylation at specific sites and gene expression in *cis*.

To first assess the efficacy and specificity of the available targeted DNA methylation editing technology, we examined the lentiviral system created by Liu et al[311] consisting of a catalytically inactive Cas9 (dCas9) fused to the catalytic domain of TET1 (dCas9-TET or a catalytic mutant, dCas9-deadTET), which is thought to promote active DNA demethylation by oxidation of the methyl moiety and eventual replacement of the modified cytosine with unmethylated cytosine by the base excision repair pathway[209].

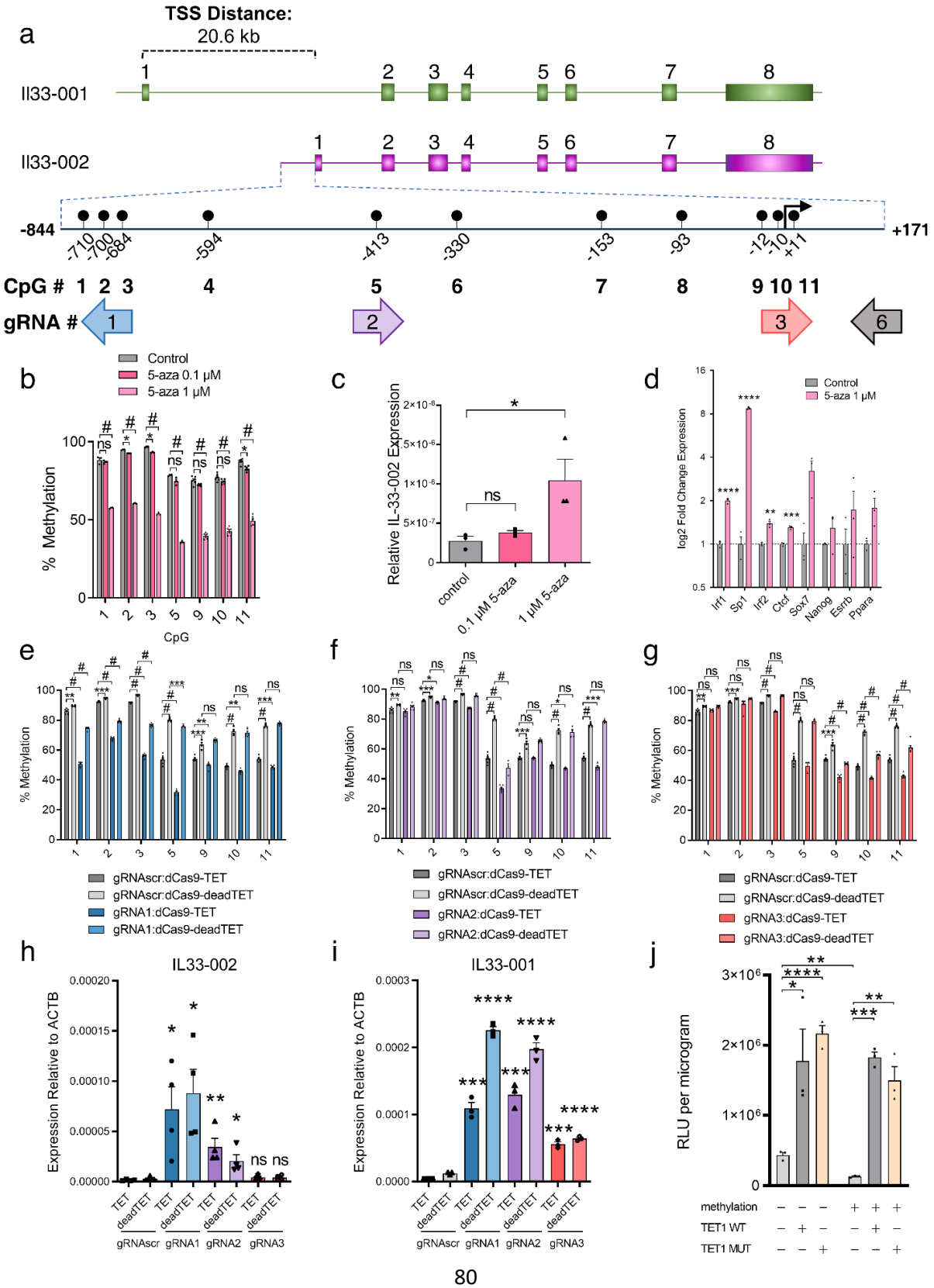

a
TSS Distance:
20.6 kb
Il33-001
Il33-002
-844
+171
-710
-700
-684
-594
-413
-330
-153
-93
-12
-10
+11
CpG # 1 2 3 4 5 6 7 8 9 10 11
gRNA # 1 2 3 6
b
Control
5-aza 0.1 μM
5-aza 1 μM
% Methylation
CpG
c
Relative IL-33-002 Expression
control
0.1 μM 5-aza
1 μM 5-aza
d
Control
5-aza 1 μM
log2 Fold Change Expression
Irf1
Sp1
Irf2
Ctcf
Sox7
Nanog
Esrrb
Ppara
e
gRNAscr:dCas9-TET
gRNAscr:dCas9-deadTET
gRNA1:dCas9-TET
gRNA1:dCas9-deadTET
f
gRNAscr:dCas9-TET
gRNAscr:dCas9-deadTET
gRNA2:dCas9-TET
gRNA2:dCas9-deadTET
g
gRNAscr:dCas9-TET
gRNAscr:dCas9-deadTET
gRNA3:dCas9-TET
gRNA3:dCas9-deadTET
h
IL33-002
Expression Relative to ACTB
TET
deadTET
gRNAscr
gRNA1
gRNA2
gRNA3
i
IL33-001
Expression Relative to ACTB
j
RLU per microgram
methylation
TET1 WT
TET1 MUT

Figure 1. Targeting the Il33 promoter with dCas9-TET. (A) Schematic of the murine *Il33* genomic locus depicting the two transcriptional isoforms with a highlighted region of an 800bp region of the *Il33-002* promoter and the locations of the 11 CpGs as well as 4 gRNAs targeting specific CpGs. The 11 CpGs are numbered sequentially in the 5' to 3' direction. The promoter-targeting gRNAs used in these experiments are shown relative to the CpGs and are approximately to scale such that CpGs 1, 2, and 3 are targeted by gRNA1, CpG 5 by gRNA 2, and gRNA 3 targets CpGs 9, 10, 11 – which overlap the transcription start site (TSS), marked by a black arrow. The orientation of the gRNAs is indicated by an arrow, where an arrow pointing to the left indicates a gRNA that binds the plus strand. The fragment cloned into the luciferase vector (pCpGl) is marked at either end, spanning from -844 to +171 relative to the TSS. (B) Percent of DNA methylation (mean ± SEM) assayed by bisulfite-pyrosequencing at the three transcription start site (TSS) CpGs (labeled 9-11) following treatment of NIH-3T3 cells with indicated concentrations of 5-aza-2'-deoxycytidine or water control (*n* = 3 independent experiments for CpGs 1, 2, 3, and 5; *n* = 6 for CpGs 9, 10, and 11) . (C) Expression of *Il33-002* (mean ± SEM) quantified by RT-qPCR and normalized to beta actin (*Actb*) expression following treatment of NIH-3T3 cells with indicated concentrations of 5-aza-2'-deoxycytidine or water control (*n* = 3 biologically independent samples) (Student's t-test, two sided, control vs. 0.1 μM 5-aza; *P* = 0.1636, control vs. 1 μM 5-aza; *P* = 0.0482). (D) Expression (mean ± SEM) of predicted (Transfac) and experimentally validated (Qiagen, ENCODE, Gene Transcription Regulation Database) *Il33-002* transcription factors quantified by RT-qPCR and normalized to *Actb* expression following treatment of NIH-3T3 cells with indicated concentrations of 5-aza-2'-deoxycytidine or water control (*n* = 3 biologically independent samples). (E-G) Percent of DNA methylation (mean ± SEM) assayed by bisulfite-pyrosequencing at 7 targeted CpGs in the *Il33-002* promoter following transduction with lentiviruses and antibiotic selection of virally infected cells (gRNAs) and selection by flow cytometry (BFP; dCas9 constructs) of NIH-3T3 cells with dCas9-Tet/dCas9-deadTET (BFP) and gRNA1 (E), gRNA2 (F), or gRNA3 (G) compared to gRNAscr (light and dark grey, gRNAscr data

identical in E-G and shown for comparison) (*n* = 4-8 biologically independent experiments, depending on specific condition and CpG; see Source Data file for specific *n* of interest). (H-I) Expression of *Il33-002* (H) and *Il33-001* (I) (mean ± SEM) quantified by RT-qPCR and normalized to *Actb* expression in NIH-3T3 stably expressing one of 4 gRNAs and dCas9-TET or dCas9-deadTET (*n* = 3-4 biologically independent samples; statistical comparisons are between each gRNA and gRNAscr bearing the same dCas9 construct (dCas9-TET or dCas9-deadTET)). All data shown as (mean+/-SEM). (J) Relative light units normalized to protein quantity (mean ± SEM) in transfected HEK293 cells. Cells were transiently transfected with methylated or unmethylated SV40-luciferase vector along with mammalian wild-type or mutant human TET1 expression plasmid or empty vector (pEF1A) control (*n* = 3). * indicates statistically significant difference of $P < 0.05$, ** of $P < 0.01$, *** of $P < 0.001$, **** or # of $P < 0.0001$, and ns = not significant (Student's t-test, two-sided, with Holm-Sidak correction if number of tests is greater than 3).

We developed a set of 20 base-pair (bp) CRISPR guide RNAs (gRNAs) targeting distinct regions in the promoter of the *Il33-002* transcript, the inducible variant[369,370] (**Figure 1A** and **Supplementary Table 1**). The system was effective in partially demethylating the *Il33* promoter; however, we noted several shortcomings of this method.

First, it was immediately apparent that even in the absence of targeting, NIH-3T3 cells expressing only a scrambled, non-targeting guide (gRNAscr) and dCas9-TET were significantly more demethylated than those expressing the same gRNAscr and dCas9-deadTET (**Figure 1E-G**). While dCas9-TET triggered a 22 to 26 percent demethylation as compared with dCas9-deadTET at CpGs 5 ($P < 0.0001$), 10 ($P < 0.0001$), and 11 ($P < 0.0001$), dCas9-TET:gRNAscr that was not targeted to these sites also caused demethylation at these sites as well as all remaining evaluated CpGs. This is indicative of a potential ubiquitous and dCas9-independent activity of the fused, over-expressed TET domain, that we provide further evidence for with whole-genome methylation analysis in a subsequent section.

Second, the demethylation caused by dCas9-TET spanned a substantial genetic distance. For example, in gRNA1:dCas9-TET cells, while the protein complex was positioned at and significantly demethylated CpGs 1, 2, and 3 ($P < 0.0001$), the remaining CpGs were all significantly demethylated as well, including CpG 11 ($P = 0.00014$), which is nearly 700 bp away from gRNA1 (**Figure 1E**). Similar significant long-distance demethylation effects could be observed in cells expressing gRNA2 and gRNA3 **(Figure 1F-G**). The potential for long-distance effects is further exemplified at the mRNA level in the strong transactivation effects of dCas9-TET positioned at the *Il33-002* promoter on the distant *Il33-001* promoter, approximately 21 kb away (**Figure 1I**).

Third, when evaluating the transcriptional effects of the epigenetic editing system, we were surprised to discover that dCas9-deadTET paired with gRNA 1 or 2 (gRNA3 blocks the TSS and likely interferes with RNA polymerase binding[304]) resulted in strong demethylation and transactivation of the *Il33-002* transcript to levels comparable to dCas9-TET (**Figure 1H**), despite lacking any catalytic capacity to initiate the active DNA demethylation process. To ensure that this unexpected result was not a consequence of erroneous sample switches, we amplified the region containing the catalytic mutations of the TET1 domain in the DNA samples used for methylation analysis and in the cDNA samples used for expression quantification and confirmed by Sanger sequencing that all dCas9-deadTET samples bore the two point mutations that render it catalytically inactive (**Supplementary Figure 1B**). Equally surprising was the fact that that dCas9-deadTET was also effective in transactivation of *Il33-001* (**Figure 1I**) ($P < 0.0001$ for all targeting gRNAs). The *Il33-001* transcript was also significantly more expressed in dCas9-deadTET cells under gRNA1 ($P = 0.0091$) and gRNA3 ($P = 0.0033$) as compared to catalytically active dCas9-TET, though it may be caused by different level of expression of the constructs; although dCas9-deadTET expression levels were moderately higher than dCas9-TET by RT-qPCR (**Supplementary Figure 1C**), the protein levels as determined by a western blot analysis were not significantly different (**Supplementary Figure 1D-E**).

To assess by a secondary measure the DNA methylation independent transactivation capacity of TET proteins, we performed transient co-transfections of *in vitro* methylated or unmethylated promoter-reporter plasmids – luciferase driven by the SV40 promoter/enhancer – in combination with a mammalian expression vector expressing human TET1 (TET1 WT), mutant TET1 (TET1 MUT) or an empty vector control (pEF1A). We found that TET1 induces the activity of completely unmethylated promoters (**Figure 1J**), as does TET2 (**Supplementary Figure 1M**), reaffirming the notion that TET proteins produce transcriptional changes independently of any DNA demethylation and thus confounding correlational assessments. SV40-pCpGl copy

number in cells is equivalent upon transfection of fully methylated or fully unmethylated DNA (**Supplementary Figure 1N**).

Additionally, we combined our three targeting gRNAs with the well-characterized dCas9-VP64 fusion; VP64 is a potent transcriptional activator originating from the herpes simplex virus[305]. The tetramer of the herpes simplex VP16 protein acts to activate transcription primarily through recruitment of basal transcription machinery, including TFIID/TFIIB, and has no known catalytic capacity for DNA demethylation [371]. Yet, we found that dCas9-VP64 co-expressed with all 3 *Il33-002* gRNAs resulted in dramatic and broad demethylation of the *Il33-002* promoter in stably infected cells (**Supplementary Figure 1F-H**). This suggests that DNA demethylation can in particular instances be secondary to transcription factor recruitment and transcriptional activation, rather than causal (**Supplementary Figure 1I**). To further test this, we performed a time-course experiment in which we observed activation of transcription of *Il33-002* by dCas9-VP64:gRNA2 24 hours after transient transfection prior to initiation of any detectable demethylation at this time point nor at any time point up to 96 hours (**Supplementary Figure 1J**). We again found significant and robust activation of the distant *Il33-001* promoter (gRNA2, $P < 0.05$; gRNA3, $P < 0.001$), supporting the notion that enzymatic domains flexibly tethered to dCas9 can act across large genetic distances (**Supplementary Figure 1K**).

Finally, we detected a significant increase of 5-hydroxymethylcytosine in the *Il33-002* promoter in the presence of dCas9-TET but not dCas9-deadTET (**Supplementary Figure 1L**), demonstrating that demethylation is not the only epigenetic change conferred by dCas9-TET and, since dCas9-deadTET activates transcription (**Supplementary Figure 1C**), that catalytic 5-hydroxymethylation is not necessary for the transcriptional induction.

A novel method for site-specific DNA methylation *in vitro*

A potential mechanism for producing specific demethylation in cells is through targeted physical interference with the DNA methyltransferase (DNMT) machinery that deposits methyl groups onto nascent post-replicative DNA. We reasoned that since dCas9 is able to interfere with transcriptional machinery to reduce gene expression[304], it may also be able to sterically obstruct DNMT activity at its binding position (**Figure 2B**). dCas9 is a prokaryotic protein with no documented protein:protein interaction with eukaryotic gene transcription machinery, the protein has no homology to known eukaryotic protein:protein interaction domains and has no enzymatic activity epigenetic or other[372].

To test this hypothesis, we first investigated whether dCas9 could be applied as a tool to interfere with DNMT activity at targeted CpGs in a simplified *in vitro* system. The target DNA used for methylation was a 1,015 bp fragment of the *Il33-002* promoter (**Figure 1A**) inserted into an otherwise CpG-free luciferase reporter vector (pCpGl)[299] to enable the assessment of methylation changes on reporter gene activity in transient transfection assays. Standard methylation with the bacterial CpG methyltransferase M.SssI protein resulted in 80-93% methylation at all CpGs as assayed by pyrosequencing (**Figure 2A**) and a significant 4-fold decrease in luciferase reporter activity in a transient transfection assay (P = 0.0041) (**Figure 2F**) . Incubation of *Il33*-pCpGl with recombinant dCas9 protein and an *in vitro* transcribed chimeric control gRNA (gRNA6 in **Figure 1A**) targeting the CpG deficient region approximately 110-130bp downstream of the TSS only slightly inhibited the efficiency of the M.SssI reaction at all CpGs (**Figure 2A** in gray). The plasmid was still highly methylated and the treatment also significantly reduced luciferase activity (P = 0.0018) compared to mock treatment and to a similar extent as standard methylation (P = 0.374) (**Figure 2F**). This confirms that the reaction components (including dCas9 protein, non-CpG-targeting gRNA, buffer system, and incubation times) do not compromise DNA methyltransferase activity.

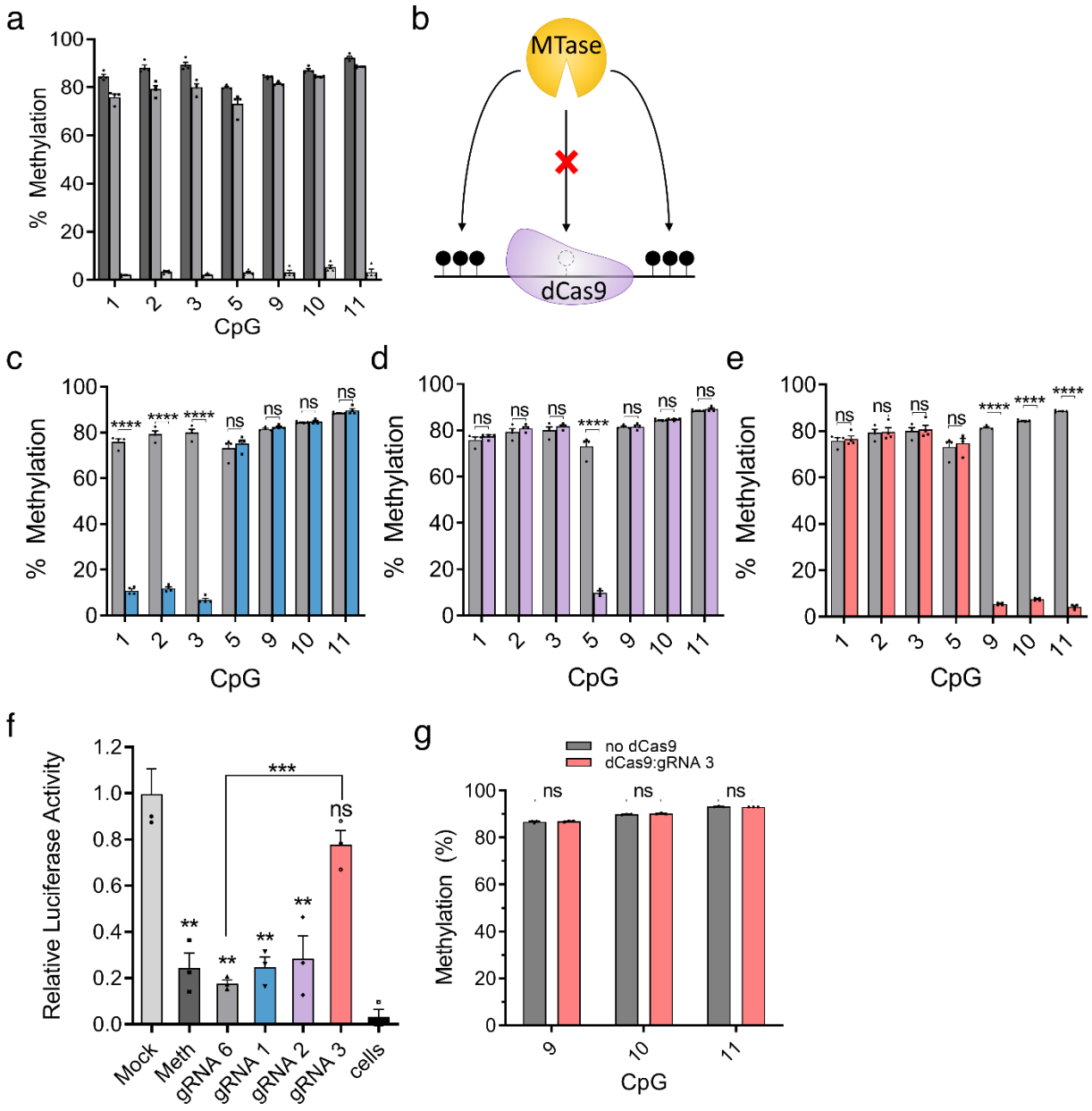


Figure 2. dCas9 blocks DNA methyltransferase *in vitro*. (A) Pyrosequencing data (mean ± SEM, *n* = 4 biologically independent samples) for the methylation state of indicated CpGs in the *Il33*-pCpGl plasmid following standard methylation for 4 hours by M.SssI (dark grey), methylation in the presence of dCas9 and gRNA 6 (distant binding) (grey), or a mock-methylated control reaction that lacked S-adenosyl methionine substrate (light grey). (B) Diagram illustrating the principle of site-specific methylation

utilizing pre-incubation of DNA with dCas9 and selective CpG-targeting guide restricting M.SssI from binding and methylating the targeted region, while permitting methylation of remaining unobstructed CpGs. (C-E) Pyrosequencing data (*n* = 4 biologically independent samples, mean ± SEM)) for the methylation state of CpGs in the IL-33-pCpGl plasmid following pre-incubation with dCas9 and gRNA1 (C), gRNA2 (D), or gRNA3 (E) and methylation by M.SssI (colored bars) . Grey bars are identical in (A, C-E) and indicate methylation levels for the same treatment utilizing gRNA6. (F) Luciferase reporter activity of the plasmids in (A, C-E), expressed as relative light units (mean ± SEM) normalized for protein content per sample, and then normalized to average value for mock methylated condition (*n* = 3 biologically independent experiments). All statistical comparisons are to mock methylated conditions unless otherwise indicated. (G) Percent of methylation (mean ± SEM) assayed by pyrosequencing when Il33-pCpGl is incubated with dCas9 and gRNA 3 or only gRNA 3 (no dCas9 control) after standard methylation, instead of before (*n* = 3 biologically independent samples). * indicates statistically significant difference of $P < 0.05$, ** of $P < 0.01$, *** of $P < 0.001$, **** of $P < 0.0001$, and ns = not significant (Student's t-test, two-sided, with Holm-Sidak correction if number of tests is greater than 3).

The DNA was then incubated with recombinant dCas9 protein and each of the three *in vitro* transcribed gRNAs – targeting CpGs in the proximal promoter region of *Il33-002* – in order to facilitate binding of the dCas9:gRNA complex to the DNA prior to the addition of M.SssI methyltransferase (**Figure 2B**). Following M.SssI treatment, the methylation state of each target CpG was assayed by bisulfite conversion and pyrosequencing and compared to treatment with control gRNA6. Pre-incubation of *Il33*-pCpGl with dCas9 and all CpG-targeting gRNAs resulted in a drastic, specific interference with DNA methylation at targeted sites (**Figure 2C-E**). For example, in the case of gRNA3, the targeted CpGs (CpGs 9, 10, and 11) were methylated only to a mean ± SEM of 5.75±0.45%, whereas the control gRNA6 barely affected methylation and the sites were methylated at 84.79±0.88% ($P < 0.00001$). Sites that were not directly within or adjacent to the binding site of dCas9:gRNA3 (CpGs 1, 2, 3, and 5) remained unaffected by the treatment (**Figure 2E**) ($P = 0.752, 0.878, 0.800, 0.618$, respectively). The same levels of inhibition and specificity were achieved by two other CpG-targeting gRNAs (**Figure 2C** and **2D**). Notably, with gRNA2, we successfully prevented methylation of a single CpG while leaving all remaining assayed CpGs completely unaffected (**Figure 2D**). We also reversed the order of the reaction, incubating the target DNA first with M.SssI and then with dCas9 and gRNA3 in order to ascertain that dCas9 is not able to catalytically remove methyl groups *post hoc* but rather inhibits methylation by competitive binding (**Figure 2G**).

Now in possession of five *Il33*-pCpGl plasmids bearing unique methylation patterns (gRNA1, gRNA2, gRNA3, gRNA6, and mock), we sought to assay the impact of these patterns on transcription in live cells using a transient transfection reporter assay. We transfected each uniquely methylated plasmid into NIH-3T3 cells and performed a luciferase reporter assay (**Figure 2F**). As mentioned previously, mock (unmethylated) plasmid drove luciferase activity to a significantly higher degree than both standard methylated and dCas9:gRNA6 treated plasmids. When CpGs 1, 2, 3 were unmethylated (by gRNA1 treatment) or CpG5 was unmethylated (by gRNA2), luciferase activity

remained low and was not significantly different from gRNA6 control (P = 0.202, P = 0.332). However, in the case of unmethylated CpGs 9, 10, and 11 (by gRNA3) surrounding the *Il33-002* TSS, luciferase activity was significantly greater than gRNA6 (P = 0.0007) and not significantly different from mock-methylated DNA (P = 0.157), demonstrating that the methylation of these three TSS CpGs, but not the others, blocks *Il33-002* promoter activity. gRNA1 and gRNA3 both interfered with methylation of 3 CpGs and thus the overall promoter methylation levels were similar between these two treatments; yet, there was a stark difference in luciferase activity. These data demonstrate the exquisite impact of site-specific methylation rather than just methylation density, and thus this assay appears to capture the sequence specificity of inhibition of promoter function by DNA methylation.

In summary, we demonstrate that dCas9 specifically inhibits DNA methylation of targeted sites *in vitro*, enabling the analysis of the causal role of specific methylated sites *per se*. The only difference between our different transfected plasmids is the positions of the methyl moieties. No additional confounding enzyme is introduced. CpGs 9, 10, and 11 at the *Il33-002* TSS silence transcription; demethylation of these CpGs is sufficient for maximal activation of the promoter-reporter construct. In contrast, demethylation of CpGs 1, 2, 3, or 5 is insufficient for re-activation of the methylated promoter suggesting that methylation of these sites is not involved in silencing of transcription from the *Il33-002* promoter.

Blocking of methylation by dCas9 is limited to its binding site and is affects both DNA strands

In the preceding *in vitro* assays, we were able to prevent on-target DNA methylation with dCas9 without affecting the remaining target CpGs in the promoter. However, as the *Il33-002* promoter is CpG-poor and clusters of CpGs (e.g. 1, 2, 3 and 9, 10, 11) are separated by several hundreds of base pairs, the precision of this approach needs to be

determined. In order to delineate the DNA span that is protected from methylation by bound dCas9, we repeated the same *in vitro* assay using a canonical CpG-rich promoter. The human *CDKN2A* (p16) promoter contains a 310 bp fragment with 38 CpGs, which are frequently aberrantly hypermethylated in all common cancers [373]. We designed a gRNA overlapping a single CpG (CpG 17) within this promoter that was flanked on either side by CpGs 8 base pairs away from the 23-nucleotide gRNA and protospacer adjacent motif (PAM) sequence (**Figure 3A**). We then applied bisulfite-cloning to map the methylation patterns of individual DNA molecules and assessed whether there was a difference in the methylation pattern of the CpGs in the strand bound by the dCas9:gRNA ribonucleoprotein and its complementary strand (as CpGs are palindromic).

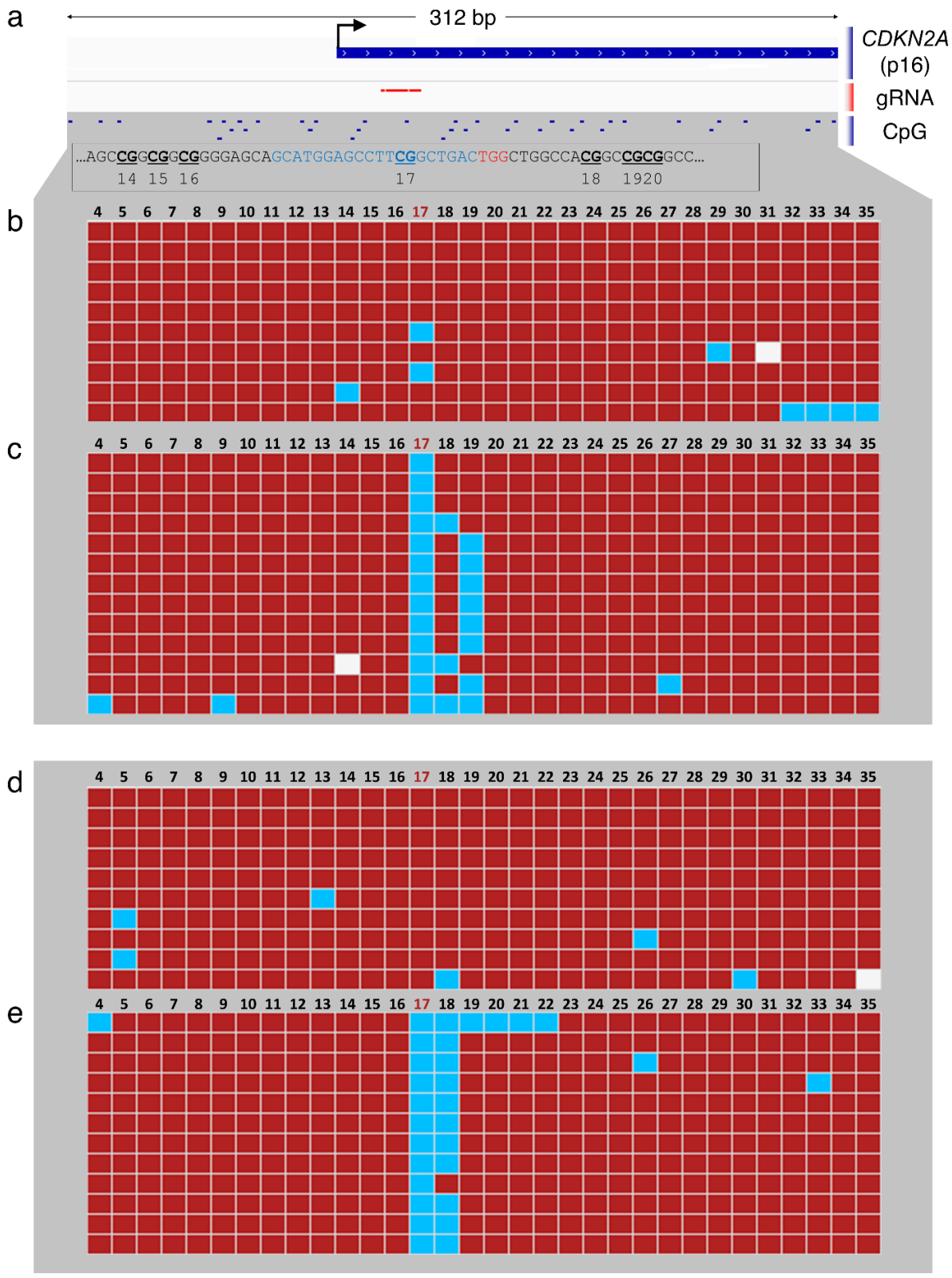

a
312 bp
CDKN2A
(p16)
gRNA
CpG
...AGCCGGCGGCGGGGAGCAGCATGGAGCCTTCGGCTGACTGGCTGGCCACGGCGCGGCCC...
14 15 16 17 18 1920
b
c
d
e
4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26 27 28 29 30 31 32 33 34 35

Figure 3. The footprint of dCas9. (A) Genome browser diagram of the *CDKN2A* (p16) promoter region, which was used for the methylation assay, showing transcription start site (TSS, marked by black arrow), gRNA position overlapping CpG 17, and surrounding CpGs. Below, DNA sequence is shown in black, gRNA sequence in blue, and PAM site in red, with CpGs bolded, underlined, and numbered according to the figures that follow. (B-E) Methylation of individual strands of the *CDKN2A* promoter plasmid following standard methylation (B,D) or methylation preceded by incubation with dCas9 and p16 gRNA (C,E). Red squares indicate methylated CpGs and blue squares indicate unmethylated CpGs; white squares indicate no data. Figures (B) and (C) represent the forward strand whereas (D) and (E) represent the reverse strand. Figures generated by BISMA software (http://services.ibc.uni-stuttgart.de/BDPC/BISMA/). Regions below 80% methylation were filtered out as strands that were not effectively methylated by M.SssI.

In the presence of a scrambled control gRNA, M.SssI almost completely methylated all CGs on both strands (**Figure 3B** and **3D**) with some sporadic unmethylated CpGs that are likely consequences of poor bisulfite conversion or Sanger sequencing errors; M.SssI is highly processive and it is unlikely that the sporadic demethylation resulted from inhibition of M.SssI[374]. In contrast, p16-targeting (CpG 17) gRNA completely inhibited methylation of the targeted CpG on the gRNA bound strand while scrambled control gRNA did not block DNA methylation of CpG 17 (0% vs. 80% methylation, $P < 0.0001$, Fisher's exact test) (**Figure 3B-C**). The CpG immediately downstream of the gRNA-PAM sequence was slightly but not significantly unmethylated (77% vs. 100% methylation, $P = 0.2292$, Fisher's exact test). Interestingly, the following CpG 19 was significantly unmethylated (38% vs. 100% methylation, $P = 0.0027$, Fisher's exact test), while the CpG only two additional base pairs downstream (CpG 20) was 100% methylated and unaffected. The distance between the unaffected CpG 20 and the 3' end of the PAM is 14 bp and the upstream unaffected CpG 16 is 8bp from the 5' end of the gRNA (**Figure 3A**). We thus define the range of dCas9 inhibition of M.SssI DNA methylation to be less than 8 base pairs from the 5' end and smaller than 14 base pairs from the 3' end of the PAM adding to a total protection range of 45 bp. Nevertheless, peak inhibition is exactly at the binding site and any inhibition within the 45 bp is only partial.

It is interesting to also note that while the target CpG 17 is always protected from methylation in all of the molecules, CpG 18 and/or CpG 19 are protected only in certain DNA molecules. These data suggest that CpGs 3' of the gRNA sequence are variably protected, possibly reflecting the dynamic orientation of the flexible gRNA scaffold [375]. It may thus be possible to refine this method to reduce or, conversely, target protection of neighboring CpGs. The results are in accordance with the crystal/cryo-EM structures of the dCas9:gRNA:DNA ternary complex, which reveals minimal 5' protrusion of dCas9:gRNA beyond the 5' end of the target DNA strand and more pronounced extension (and steric interference) of both dCas9 protein and gRNA scaffold beyond the

3’ end of the target DNA sequence, which still seats deep within the dCas9 binding pocket (**Supplementary Figure 2A**)[375,376].

We also determined whether protection from methylation by dCas9 was symmetric on both DNA strands and whether dCas9 preferably obstructed methylation of the targeted CpG only on the strand that was complementary to the gRNA. Given that bound dCas9 envelopes nearly the entire DNA double helix [375], we predicted that both CpG sites would be equally protected. Bisulfite-cloning of the opposite strand again revealed complete protection from M.SssI methylation of CpG 17 (0% vs. 100%, $P < 0.0001$) and the next CpG (8% vs. 90%, $P = 0.0003$) (**Figure 3D-E**). Interesting, the 3’ footprint is smaller by at least 2bp (and at most 6bp) than in the strand interacting with the gRNA, as CpG 19 is not affected on the antisense strand. Thus, dCas9:gRNA complex completely protected both the target and complementary CpG on the antisense strand. We determined whether we could focus the range of protection using the smaller dCas9 protein from *Staphylococcus aureus* despite the fact that it requires a longer gRNA (21bp instead of 20bp) and a longer PAM sequence (NNGRR instead of NGG). We designed 4 *S. aureus* gRNAs (SAgRNAs1-4) that also overlapped with potential gRNAs for the hitherto utilized *Streptococcus pyogenes* dCas9 (SPgRNAs1-4) (**Supplementary Figure 2B**). The first three gRNAs assayed the 5’ protrusion and were shifted by one base pair each in order to refine the 5’ distance for both dCas9 variants; three 5’ CpGs were 4, 7, and 11 bp away from the 5’ end of SAgRNA1; 3, 6, and 10 bp away for SAgRNA2; and 2, 5, and 9 bp away for SAgRNA3. Each CpG was 1 bp further away for the corresponding SPgRNAs as these were 1 bp shorter at the 5’ end (20 bp vs. 21 bp). We determined that *S. aureus* dCas9 is equally capable of complete interference with M.SssI at sites within the bound region (CpGs 20-22), with a gradual 5’ fall-off in protection; 90%-100% protection of CpG 2-4bp away, 80% protection of CpG 5 bp away, 50-60% at CpG 6 or 7 bp away and 0-10% at 9-11bp away from the target (**Supplementary Figure 2C** and **S2D**). 5’ interference of SP-dCas9 was consistently less than SA-dCas9 at all distances in a manner that was not sufficiently explained by

the additional single 5' bp of the *S. aureus* gRNAs (**Supplementary Figure 2D**). The 3' distance for SP-dCas9 could not be refined further because of a lack of efficacy of SPgRNA 4 (**Supplementary Figure 2C** and **E**); only 4 strands appeared to have been protected from dCas9 (of 17 sequenced) and the interference was interestingly limited to CpGs in the PAM site and not within the gRNA binding site, likely indicative of a poor-quality gRNA. However, SAgRNA4 was efficient and we could calculate that SAdCas9 interfered with a minimum of 11bp and a maximum of 13bp from the 3' end, including its 5bp PAM sequence. Therefore, we demonstrate that despite its smaller protein size, SA-dCas9 has a 3' footprint comparable to but possibly smaller than SP-dCas9 (likely due to similar gRNA scaffolds) and a definitively larger 5' footprint, drawing the conclusion that the original SP-dCas9 allows more precise interference with DNMTs, however it is also useful to note that the equivalent efficacy of SA-dCas9 presents a secondary option for combinational approaches and for a more diverse selection of target sequences by addition of a second PAM option.

The dCas9 system directs robust site-specific demethylation in living cells

dCas9 is obviously not an active demethylase; nevertheless, we hypothesized that we could use it to demethylate specific CpGs in living dividing cells. As nascent post-replicative DNA is unmodified and must be methylated by the maintenance methyltransferase Dnmt1 in order to preserve parental cell methylation patterns [377], we postulated that dCas9 would interfere with Dnmt1 methylation similar to its blockage of M.SssI methylation and thereby cause passive demethylation of targeted sites through successive rounds of cell division and DNA replication. Therefore, we used the gRNAs characterized above to demethylate the endogenous *Il33-002* promoter in NIH-3T3 cells. We established by lentiviral transduction cell lines stably expressing SP-dCas9 and each *Il33* gRNA or a scrambled, non-targeting control gRNA (gRNAscr) and collected DNA for methylation analysis by bisulfite conversion and pyrosequencing one week after complete antibiotic marker selection. We demonstrate that the dCas9:gRNA

complex is sufficient to produce robust demethylation of targeted CpGs (**Figure 4A-C**). dCas9 in combination with gRNA1 (**Figure 4A**) reduced absolute methylation levels by an average of 27.0% ($P < 0.0001$), 28.3% ($P < 0.0001$), and 34.6% ($P < 0.0001$) at CpGs 1, 2, and 3, respectively; gRNA2 (**Figure 4B**) reduced CpG 5 methylation by 52.0% ($P < 0.0001$); gRNA3 (**Figure 4C**) reduced CpG 9, 10, and 11 methylation by 30.2%, 31.4%, and 38.4% ($P < 0.0001$ for all). Demethylation with dCas9, unlike dCas9-TET (**Figure 1F**) was highly specific to targeted CpGs, as in the case of gRNA2, no other assayed CpGs were demethylated. gRNA3 caused significant demethylation of off-target CpG 3 ($P = 0.002$) but the extent of demethylation was only 0.6%. gRNA1 caused a slightly larger, significant demethylation of the distant CpGs 9, 10, and 11 (5.3%, 4.5%, and 3.9%) but still to a lower level than the target CpGs 1, 2, and 3, and less than that of dCas9-TET:gRNA1. These data also clarify that the binding site demethylation in dCas9-TET and in dCas9-deadTET cells (**Figure 1E-G**) likely stems from the same mechanism of steric interference with Dnmt1 rather than a catalytic TET activity, as the tightly bound dCas9 domain likely makes it impossible for the fused TET domain to access this bound DNA.

We were also able to demonstrate similar levels of demethylation and specificity by a second gRNA targeting CpGs 9, 10, and 11 which was shifted two base pairs in the 3' direction (**Supplementary Figure 3A**) relative to gRNA3, demonstrating that altering the exact CpG positioning relative to the gRNA, whether within the gRNA target sequence, PAM site, or immediately adjacent to either, does not impact demethylation efficiency in cells. All these positions were predicted to be completely protected from DNMT activity by both gRNAs in the *in vitro* footprint assays (**Figure 3**).

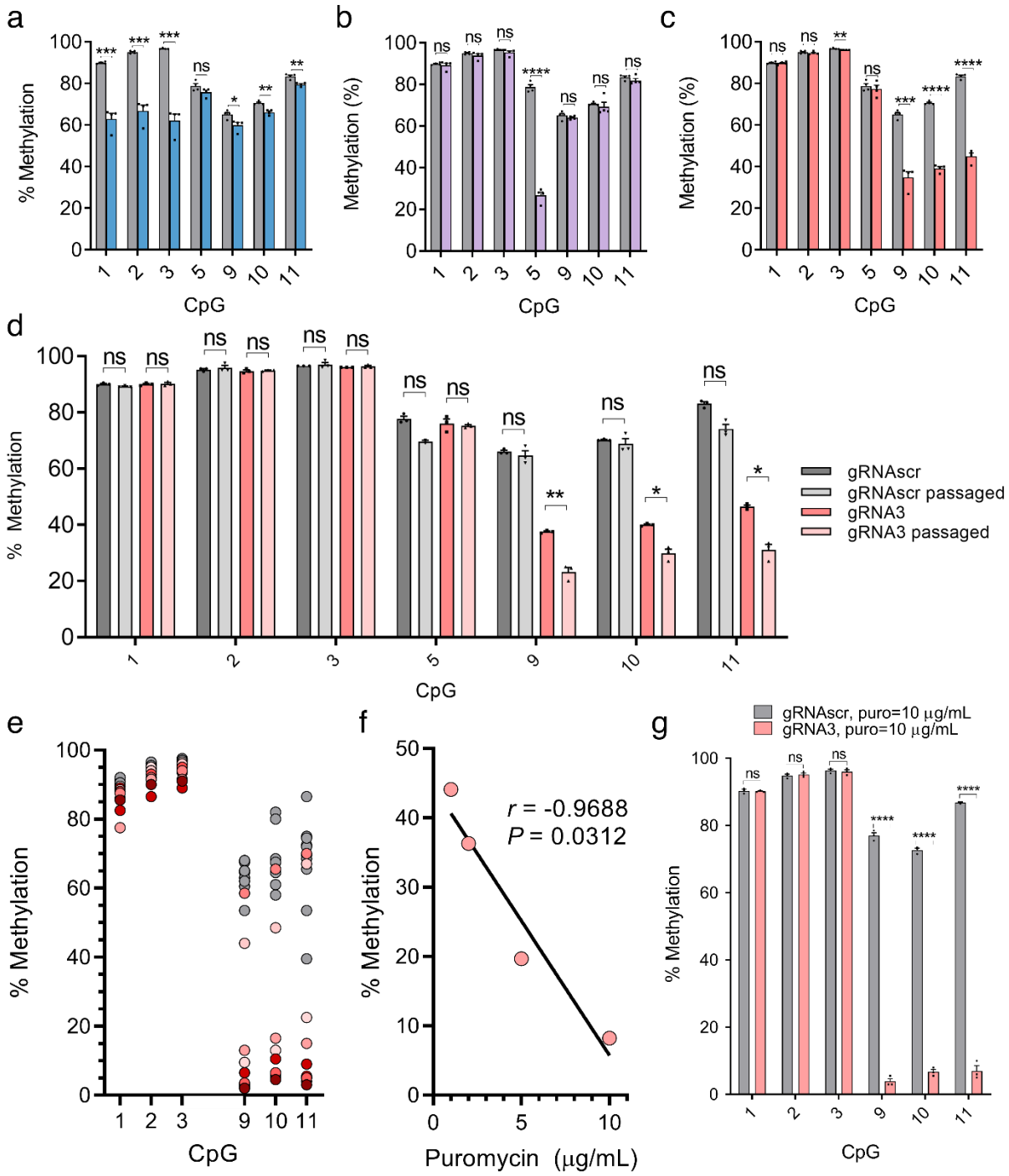

a
% Methylation
CpG
b
Methylation (%)
CpG
c
Methylation (%)
CpG
d
% Methylation
CpG
gRNAscr
gRNAscr passaged
gRNA3
gRNA3 passaged
e
% Methylation
CpG
f
% Methylation
r = -0.9688
P = 0.0312
Puromycin (μg/mL)
g
gRNAscr, puro=10 μg/mL
gRNA3, puro=10 μg/mL
% Methylation
CpG

Figure 4. dCas9 causes demethylation in mammalian cells. (A-C) Methylation levels (mean ± SEM) assayed by bisulfite-pyrosequencing at CpGs 1, 2, 3, 5, 9, 10, and 11 of NIH-3T3 cells stably expressing dCas9 and gRNA1 (A, blue), gRNA2 (B, purple), gRNA3 (C, pink) or scrambled gRNA (A-C, grey; identical in all) (*n* = 4 biologically independent samples). (D). Cells from (C) were passaged for an additional 30 days and methylation percentage was assayed as previously (*n* = 3 biologically independent samples, mean ± SEM). (E) Cells from (C) were subjected to clonal isolation and expansion. Grey circles represent methylation levels of clones containing dCas9 and scrambled gRNA and various red circles represent methylation levels of randomly selected clones stably expressing dCas9 and gRNA 3 (*n* = 10 independent clones per condition). (F) Average DNA methylation at CpGs 9-11, assayed by bisulfite-pyrosequencing, as a function of increasing the selection antibiotic puromycin (lentivirus is expressing puromycin resistance gene) concentration in cell lines (pools) stably expressing dCas9 and gRNA3 (*n* = 1 cell line per puromycin concentration) fitted with a line of best fit. (G). DNA methylation at CpGs 1, 2, 3, 9, 10, and 11 in (*n* = 3 biologically independent samples, mean ± SEM) NIH-3T3 cells stably expressing dCas9 and gRNA3 (pink) or control gRNAscr (grey) and treated with 10 μg/mL puromycin until no antibiotic-associated cell death could be observed and surviving cells were of sufficient quantity for DNA extraction and other procedures (approximately 2 weeks). * indicates statistically significant difference of $P < 0.05$, ** of $P < 0.01$, *** of $P < 0.001$, **** of $P < 0.0001$, and ns = not significant (Student's t-test, two-sided, with Holm-Sidak correction if number of tests is greater than 3).

Though these experiments demonstrated a higher specificity of dCas9 than dCas9-TET across adjacent CpGs in the *Il33-002* promoter, we also sought to determine if the same off-target effects seen with dCas9-TET could be found in equivalent dCas9 treated cells. Unlike in dCas9-TET cells, the distant *Il33-001* transcript was not upregulated by dCas9 combined with any of the three targeting gRNAs (**Supplementary Figure 3B**); however, there was detectable significant downregulation of *Il33-001* under gRNA1. We found no potential off-target site for gRNA1 (no less than 8 mismatches) within +/- kb from the *Il33-001* TSS.

Next, we wished to evaluate if the dCas9 demethylation approach could be optimized to yield higher demethylation. Passive demethylation by Dnmt1 interference would require cell division and if fully efficient, methylation levels would halve with every round of replication. We therefore hypothesized that passaging the cells in culture would increase the extent of demethylation. dCas9:gRNA3 and dCas9:gRNAscr cell lines were passaged for an additional 30 days after the original DNA collection. This approach increased the extent of demethylation of only CpGs 9 (14.3%, $P = 0.0009$), 10 (10.2%, $P = 0.003$), and 11 (15.5%, $P = 0.002$) (**Figure 4D**). Passaged dCas9:gRNAscr cell lines were demethylated at several CpGs compared to original unpassaged cells but none of these differences were significant after correction for multiple testing.

Another common approach to improve the efficiency of CRISPR/Cas9 editing is cloning [378]. Despite the fact that we could achieve robust demethylation of a target CpG in a population of cells, as a particular strand of DNA only exists in a methylated or unmethylated state, we reasoned that we could isolate clonal populations that are completely demethylated at the target sites (CpG 9,10,11). Therefore, we expanded 10 clonal lines from each of the dCas9:gRNA3 and dCas9:gRNAscr cell lines and subjected these clones to pyrosequencing. The population of gRNAscr clones (grey circles) was not significantly demethylated relative to the original gRNAscr pool at any CpG except a significant 0.6% demethylation at CpG 3, and, with the lone exception of

a single CpG in one clone that displayed 39.5% methylation, no CpG in any of the 10 clones was methylated less than 50% (**Figure 4E**). Therefore, even though some gRNAscr cells in a population that is not 100% methylated must have fully unmethylated CpGs, the clonal isolation process is unable to generate fully demethylated clones, perhaps due to a given equilibrium between methylation and demethylation established by the nuclear DNA methylation machinery in the cells. dCas9:gRNA3 clones were not significantly demethylated at target CpGs 9, 10, and 11, compared to both original and passaged lines. However, 6 of 10 clones isolated from the dCas9:gRNA3 pool displayed methylation levels below 11% at CpGs 9, 10, and 11 and two of these clones were methylated at or below 5% at all targeted CpGs. We concluded that we were able to produce cell lines with almost completely demethylated target CpGs with this approach (the small level of methylation detected in these clones is around the standard error for unmethylated controls in our pyrosequencing assay).

The clonal analysis suggests a clonal variation in the extent of demethylation by dCas9:gRNA. A plausible cause could be variation in the level of expression of either dCas9 or the gRNA. dCas9 mRNA levels did not correlate with methylation levels (r = 0.1982, P = 0.6091, n = 9) (**Supplementary Figure 3D**) whereas gRNA3 expression levels correlated negatively with methylation (r = -0.7307, P < 0.05, n = 9) (**Supplementary Figure 3E**). Similar to several other studies that demonstrated that expression of gRNA is the rate limiting factor in Cas9 cleavage efficiency [379-382], our data suggest that gRNA is the limiting factor in targeted demethylation efficiency.

Clonal isolation is tedious, involves long passaging times, and prone to producing bottleneck effects from a heterogenous cell line; we also found that unhealthy morphologies were common to these clonal populations (**Supplementary Figure 4**). In order to increase gRNA transgene expression in the clonal population, we increased the quantity of puromycin, which we hypothesized would select for cells with higher copy numbers of virally-inserted transgenes and increased output of gRNA expression. We

noted a stepwise increase in demethylation as puromycin concentrations were increased from the standard 1 µL/mL concentration to 2, 5, or 10 µg/mL (**Figure 4F**) with a significant correlation ($P < 0.05$) and a large difference of 36% in extent of demethylation of the target sequences between minimal and maximal concentrations. Settling on 10 µg/mL, we produced high-puromycin selected populations of gRNAs1-3 and gRNAscr and verified the extent of demethylation. We found that dCas9:gRNA3-treated cells were highly demethylated at CpGs 9-11 with 3-10% residual methylation, compared to 71-87% in dCas9:gRNAscr cells with 10 µg/mL puromycin ($P < 0.000001$ for all), while off-target CpGs 1-3 were still highly methylated and unaffected by the treatment ($P = 0.742$, 0.621, and 0.670, respectively) (**Figure 4G**). In summary, we successfully developed a protocol to produce near-complete, specific targeted DNA demethylation in cell lines and selected this optimized approach for future experiments.

The effect of site-specific demethylation on *Il33* gene expression

The next step was to assess the utility of our demethylation strategy in exploring the causal links between DNA demethylation at a specific region and transcriptional changes. We predicted that demethylation in this context would not be sufficient to activate transcription because dCas9 remains bound to the TSS and obstructs binding of transcriptional machinery, which is in itself an established technique to inhibit gene expression [304]. Accordingly, despite robust demethylation, high-puromycin dCas9:gRNA3 cell lines expressed significantly less *Il33-002* transcript than even scrambled cells (**Supplementary Figure 3F**) whereas the *Il33-001* isoform was not significantly impacted in the same cells (**Supplementary Figure 3B**). In fact, in contrast to the typical negative correlation between expression and DNA methylation, *Il33-002* expression was positively correlated with CpG 9-11 methylation level across dCas9:gRNA3 clones ($r = 0.74$, $P = 0.02$) (**Supplementary Figure 3G**). This unique relationship likely originates from the fact that increased dCas9 on-target binding not

only obstructs Dnmt1 activity but also concurrently blocks access to RNApolII complex, inhibiting transcription.

To study the transcriptional consequences of promoter demethylation, dCas9 would need to be removed following demethylation to expose the newly unmethylated DNA to the nuclear environment. We tested transient gRNA expression with the aim that following several rounds of cell division, having caused demethylation of target DNA, gRNAs will be diluted and will not block binding of RNApolII. However, transient transfection of guide RNA molecules in a stably expressed dCas9 background resulted in only 15% on-target demethylation (**Supplementary Figure 5**) and we determined to forego optimization of this strategy in favor of one compatible with the optimized high-puromycin protocol we had established. We implemented the Cre-lox system (**Figure 5A**) that would allow complete dCas9 removal by Cre-recombination only after demethylation is maximized.

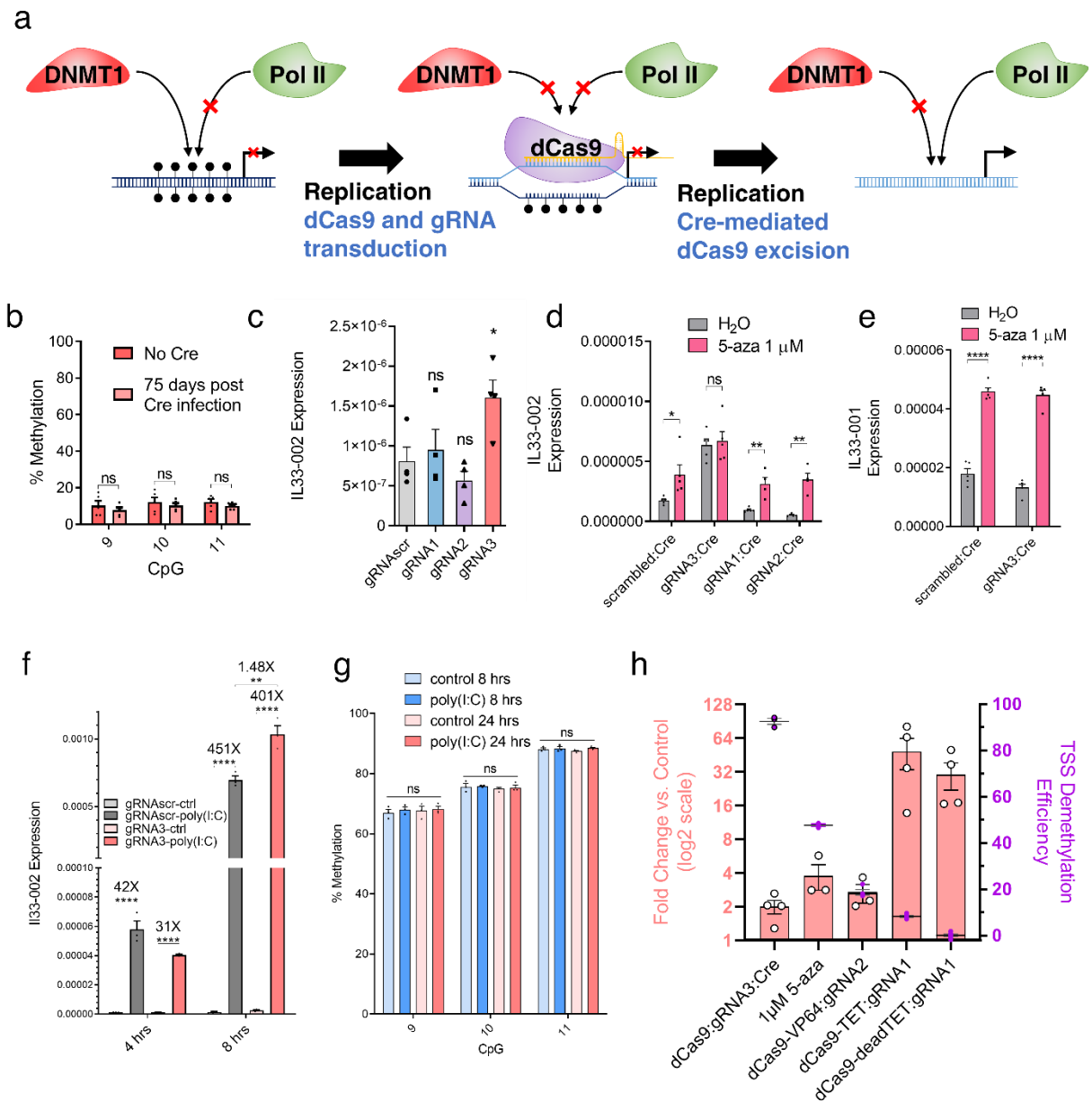

a
DNMT1
Pol II
dCas9
Replication
dCas9 and gRNA
transduction
Replication
Cre-mediated
dCas9 excision
b
% Methylation
No Cre
75 days post
Cre infection
ns
CpG
c
IL33-002 Expression
gRNAscr
gRNA1
gRNA2
gRNA3
d
IL33-002
Expression
H2O
5-aza 1 μM
scrambled:Cre
gRNA3:Cre
gRNA1:Cre
gRNA2:Cre
e
IL33-001
Expression
scrambled:Cre
gRNA3:Cre
f
Il33-002 Expression
gRNAscr-ctrl
gRNAscr-poly(I:C)
gRNA3-ctrl
gRNA3-poly(I:C)
1.48X
401X
451X
42X
31X
4 hrs
8 hrs
g
control 8 hrs
poly(I:C) 8 hrs
control 24 hrs
poly(I:C) 24 hrs
% Methylation
CpG
h
Fold Change vs. Control
(log2 scale)
TSS Demethylation
Efficiency
dCas9:gRNA3:Cre
1μM 5-aza
dCas9-VP64:gRNA2
dCas9-TET:gRNA1
dCas9-deadTET:gRNA1

Figure 5. The effect of targeted promoter DNA demethylation on *Il33* expression. (A) Diagram illustrating the principle of site-specific demethylation with dCas9 removal in order to facilitate transcription factor binding to the newly demethylated region. First, DNA is endogenously methylated by Dnmt1 with every round of replication and RNA polymerase II (RNA-polII)is not recruited to the promoter. After the introduction of dCas9 and a promoter-targeting gRNA, Dnmt1 is physically occluded from the locus and nascent strands of DNA are unmethylated, facilitating passive demethylation of the bound region. However, RNA-polII is also physically occluded by dCas9. If dCas9 is successfully removed, the unmethylated DNA no longer serves as a substrate for Dnmt1 and continues to remain unmethylated and RNA-polII may now be recruited. (B) Methylation of CpGs 9, 10, and 11 (mean ± SEM) which had been previously demethylated by high-puromycin gRNA3:dCas9 in NIH-3T3 cells, after 75 days of passaging following the lentiviral transduction of Cre recombinase (pink) or empty-vector control (red) (*n* = 5 biologically independent samples). (C) *Il33-002* expression (mean ± SEM) in NIH-3T3 cell lines stably expressing gRNAscr (grey) or gRNA1 (blue), gRNA2 (purple), or gRNA3 (pink) under high-puromycin conditions in combination with dCas9, followed by dCas9 removal by Cre recombinase as assayed by RT-qPCR and normalized to *Actb* expression. Statistical comparisons are to gRNAscr condition (*n* = 4 biologically independent samples). (D-E) *Il33-002* expression (D, *n* = 4-5 biologically independent samples) or *Il33-001* expression (E, *n* = 5 biologically independent samples) in NIH-3T3 cells from (C) following treatment wither water control or 1 μM 5-aza-2'-deoxycytidine, measured by RT-qPCR and normalized to *Actb* expression (mean ± SEM). (F) *Il33-002* expression (mean ± SEM) measured by RT-qPCR and normalized to *Actb* expression, in dCas9:gRNAscr (grey) or dCas9:gRNA3 (pink) NIH-3T3 cells following Cre recombinase treatment and then treated with poly(I:C) (1 μg/mL) or water control for 4 or 8 hours (*n* = 3 biologically independent experiments). (G) DNA methylation (mean ± SEM) assayed by bisulfite-pyrosequencing in NIH-3T3 cells expressing dCas9, gRNAscr, and Cre treated with 1 μg/mL poly(I:C) or water control for 8 hrs and 24 hrs (*n* = 3 biologically independent experiments). (H) Summary of maximal

Il33-002 induction (mean ± SEM) (left y-axis, pink bars; data in log2 scale but axis numbering is not transformed) and maximal promoter demethylation (purple, right y-axis, calculated as percent unmethylated divided by control methylation) under different treatments presented thus far (x-axis: dCas9, 5-aza-2'-deoxycytidine, dCas9-VP64, dCas9-TET, and dCas9-deadTET). Where relevant, data for maximally inducing/demethylating gRNA is shown. n = 3-6 biologically independent cell cultures * indicates statistically significant difference of $P < 0.05$, ** of $P < 0.01$, *** of $P < 0.001$, **** of $P < 0.0001$, and ns = not significant (Student's t-test, two-sided, with Holm-Sidak correction if number of tests is greater than 3).

We established new high-puromycin selected NIH-3T3 cell lines expressing each lentiviral *Il33* gRNA and a lentiviral loxP-flanked dCas9 variant and validated successful demethylation (**Supplementary Figure 6A-C**). One of the two base substitutions to render this dCas9 variant nuclease-dead (D10A, H840A) is different than the dCas9 used in previous experiments (D10A, N863A). We then used lentivirus-mediated gene transfer to introduce either Cre recombinase or an empty control vector and verified successful dCas9 removal by Cre at the DNA level by PCR, using primers that produced a 500bp fragment upon recombination (**Supplementary Figure 6D-E**), and at the protein level by chromatin immunoprecipitation followed by quantitative PCR (ChIP-qPCR). ChIP-qPCR demonstrated elevated dCas9 binding to the *Il33-002* promoter region only in cells stably expressing dCas9 and gRNA3 but not in dCas9:gRNAscr cells regardless of Cre treatment (**Supplementary Figure 6F**); in dCas9:gRNA3 cells, Cre recombination eliminated dCas9 binding to the *Il33-002* promoter to levels with no significant difference from dCas9:gRNAscr cells. Interestingly, low levels of methylation persisted for at least 75 days after removal of dCas9 by Cre recombinase (**Figure 5B**), indicating a lack of *de novo* methylation of this locus in these cells and the ability of this approach to modify DNA methylation in a stable manner despite elimination of dCas9.

Having generated targeted demethylation without bound dCas9 to hinder RNApolII binding to the TSS, we were then able to interrogate whether demethylation of the proximal promoter causes changes in expression of the gene. Expression levels of *Il33-002* transcript were measured by RT-qPCR. We detected a small but significant (P = 0.0312) increase in expression in NIH-3T3 cells treated with dCas9:gRNA3 and Cre recombinase as compared to dCas9:gRNAscr, but not in dCas9:gRNA1 or dCas9:gRNA2 cells (**Figure 5C**). This is consistent with our *in vitro*/transient transfection luciferase assays findings (**Figure 2F**); both approaches suggest that methylation of TSS CpGs 9, 10, and 11 silence the basal *Il33-002* promoter.
It is possible that the small magnitude of induction of expression by demethylation of the TSS region can be explained by the presence of other methylated regulatory regions or

other required *trans*-acting factors that need to be demethylated to facilitate larger changes in expression. We used 5-aza-2'-deoxycytidine, a global demethylation agent, to assess whether demethylation of other sites would further induce the expression of TSS-demethylated *Il33-002*. Our results show that gRNAscr-, gRNA1-, and gRNA2-bearing cells, which were still methylated at the TSS, were still induced by the drug, while gRNA3 treated cells that were demethylated at the TSS were no longer responsive (Fig 5D.), suggesting that no further demethylation is required beyond demethylation of TSS sites 9, 10, and 11 for the activity of the basal promoter. To further corroborate that the lack of further induction by 5-aza-2'-deoxycytidine in cells with demethylated CpG sites 9, 10, and 11 was not a consequence of some other resistance to demethylation of dCas9:gRNA3 cells, we demonstrate that, in these dCas9:gRNA3 cells, the induction of the *Il33-001* isoform, driven by an untargeted upstream promoter, continued to be responsive to 5-aza-2'-deoxycytidine (**Figure 5E**).

We verified that lack of further induction of gRNA3 demethylated *Il33-002* by a demethylating agent was not a result of an upper threshold of expression or our detection method, because treatment of cells with 1 µg/mL polyinosinic:polycytidylic acid (poly(I:C)) activated expression of *Il33-002* several hundred-fold after 4 and 8 hrs (**Figure 5F**). Equally surprising was the fact that that dCas9:gRNA3 induced a 1.48X higher level of *Il33-002* expression than dCas9:gRNAscr counterparts at 8 hours (P = 0.0097). However, the overall induction within each treatment group (poly(I:C) vs. control) was lower in gRNA3 cells (401X) than in gRNAscr cells (451X), because control-treated gRNA3 cells already have a higher baseline *Il33-002* expression as demonstrated here (1.67X, P = 0.1354) and in **Figure 5C-D**. Interestingly, this strong induction in response to poly(I:C) occurred in the complete absence of any detectable demethylation of the three TSS CpGs after 8 hours (in gRNAscr cells) and even when incubation was extended to 24 hours (**Figure 5G**) nor of any other CpGs in the promoter (**Supplementary Figure 7A**). These data suggest that DNA methylation

suppresses basal activity of the *Il33-002* promoter but does not dramatically affect its inducibility, which can be independent of DNA methylation in the promoter region.

Histone deacetylase inhibition has been previously reported to act in combination with DNA demethylation to activate gene expression[383]. Activation of gene expression might require both demethylation and histone acetylation. We tested whether we can achieve a robust activation of the demethylated *Il33-002* with the histone deacetylase inhibitor trichostatin A (TSA, 50 nM). However, we only noticed a minor difference in the responses to treatment with TSA: in gRNAscr, gRNA1, and gRNA2 cells, TSA slightly reduced expression, and in gRNA3 cells, expression was not affected by TSA (**Supplementary Figure 7B**). Thus, TSA inhibition of histone deacetylase activity does not add to the transcription activity of *Il33-002*. Finally, we determined whether demethylation poises *Il33-002* to activation by other inducers. Lipopolysaccharide (LPS) has previously been shown to induce *Il33* [384]. In a pattern nearly identical to poly(I:C), treatment of NIH-3T3 cells with 100ng/mL LPS induced the overall expression levels of *Il33-002* in gRNA3 cells where the TSS is demethylated to a larger extent (2.03X, $P < 0.01$) than in cells were *Il33-002* TSS is methylated (gRNAscr), however the fold change within each condition relative to phosphate-buffered saline (PBS) control was similar (2.43X in gRNA3 and 2.74X in gRNAscr) as a consequence of 2.28X higher baseline *Il33-002* expression in gRNA3 cells treated with PBS than gRNAscr treated with PBS, consistent with our observations in **Figure 5** (**Supplementary Figure 7C**). This suggests that LPS can activate both the unmethylated and methylated *Il33-002*, but the total output increases once the promoter is demethylated. Alternatively, since *Il33* is not 100% methylated in control cells and in some cells the promoter is unmethylated (~20%), LPS might have induced the unmethylated copies in the control cells explaining the lower total output in the control cells. However, the ratio of unmethylated Il33 promoter (20%) in the untreated cells relative to the demethylated cells (90%) (0.22) is lower than the ratio of expression in control and demethylated cells following LPS

induction (0.5). The data are consistent with the hypothesis that at this locus methylation can silence basal promoter activity but not affect inducibility.

In summary, we show that near-complete demethylation of *Il33-002* TSS using an enzyme-free approach results in only a mild two-fold induction of basal gene expression, whereas other approaches that cause smaller degrees of demethylation can produce larger changes in gene expression, such as dCas9-TET:gRNA1, which produces only a 10% demethylation but a 50-fold gene induction (**Figure 5H**).

dCas9 off-target demethylation events and comparison to dCas9-TET

In order to determine the specificity of dCas9-targeted demethylation and compare it to that of the dCas9-TET method, we performed whole-genome bisulfite sequencing (WGBS) of control (untreated) NIH-3T3 cells, dCas9-TET:gRNA3 and dCas9-TET:gRNAscr NIH-3T3 cells (from **Figure 1**), as well as dCas9:gRNA3 and dCas9:gRNAscr NIH-3T3 cells subsequently treated with lentiviral Cre recombinase (Cre rationale provided in the following section) (n = 3). To understand the changes imposed by these treatments at a global level, we first performed an analysis of CpG methylation clustering of high-coverage (≥10X) CpGs genome-wide, from which it was apparent that cells modified with the dCas9 method clustered with untreated control cells, whereas dCas9-TET cells were (1) more divergent from control cells and (2) unable to be clustered within gRNAscr vs gRNA3, reaffirming the global effects of TET despite targeting by dCas9:gRNA (**Figure 6A**). dCas9-TET cells with both gRNAs were also significantly less methylated genome-wide than untreated cells (P < 0.01) and dCas9 counterparts (P < 0.01 for gRNAscr, P < 0.01 for gRNA3) to such an extent that they failed to demonstrate the typical genomic hypermethylation in response to lentiviral integration [385,386] that dCas9 cells demonstrated (**Figure 6B**).

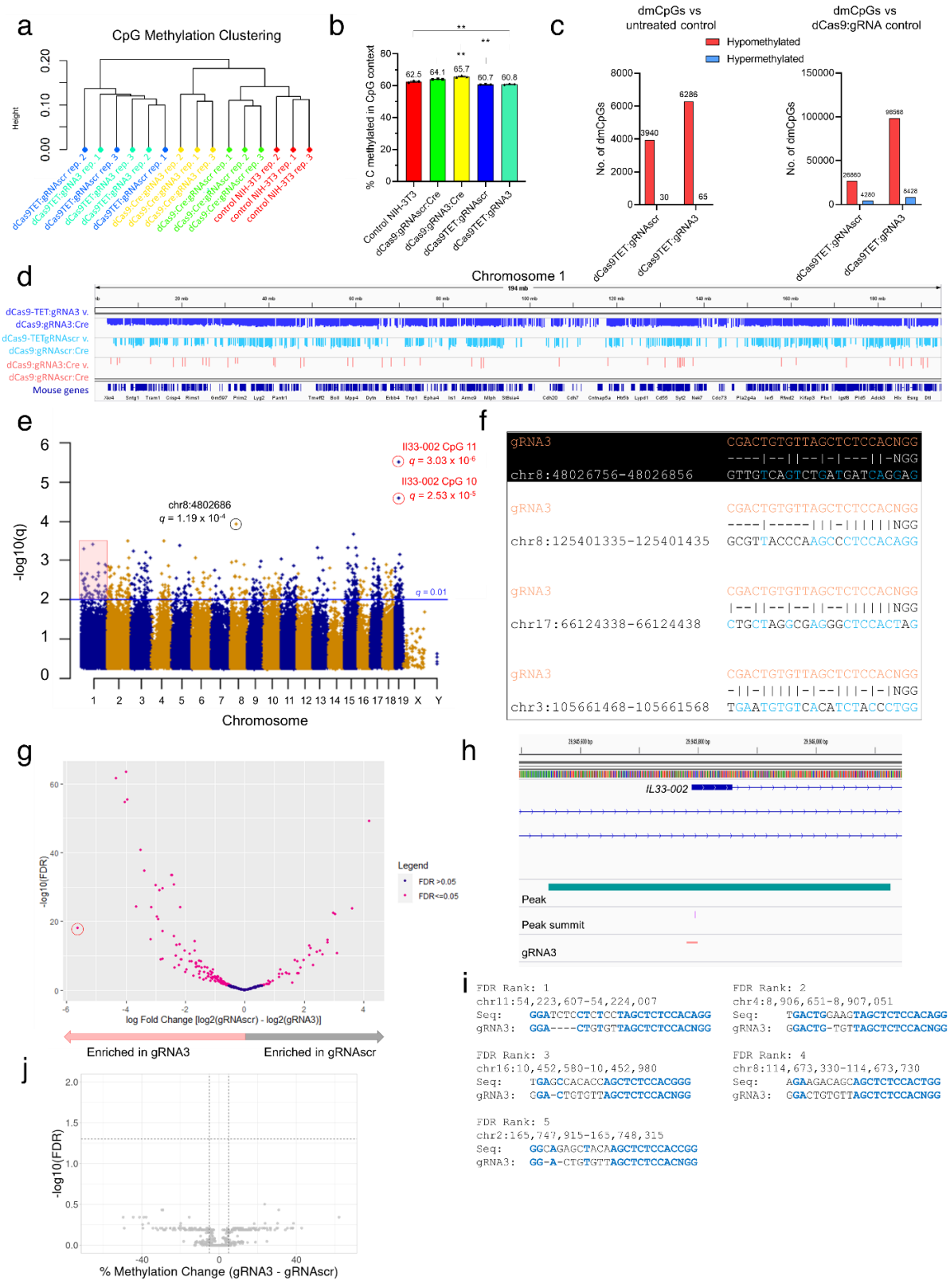

a
CpG Methylation Clustering
Height
0.20
0.10
0.00
dCas9TET:gRNAScr rep. 2
dCas9TET:gRNA3 rep. 1
dCas9TET:gRNAScr rep. 3
dCas9TET:gRNA3 rep. 3
dCas9TET:gRNAScr rep. 1
dCas9TET:gRNA3 rep. 2
dCas9:Cre:gRNA3 rep. 1
dCas9:Cre:gRNAScr rep. 1
dCas9:Cre:gRNA3 rep. 2
dCas9:Cre:gRNAScr rep. 2
dCas9:Cre:gRNA3 rep. 3
dCas9:Cre:gRNAScr rep. 3
control NIH-3T3 rep. 1
control NIH-3T3 rep. 2
control NIH-3T3 rep. 3
b
% C methylated in CpG context
62.5
64.1
65.7
60.7
60.8
Control NIH-3T3
dCas9:gRNAScr:Cre
dCas9:gRNA3:Cre
dCas9TET:gRNAScr
dCas9TET:gRNA3
c
dmCpGs vs untreated control
Hypomethylated
Hypermethylated
No. of dmCpGs
3940
30
6286
65
dCas9TET:gRNAScr
dCas9TET:gRNA3
dmCpGs vs dCas9:gRNA control
No. of dmCpGs
26860
4280
98568
8428
dCas9TET:gRNAScr
dCas9TET:gRNA3
d
Chromosome 1
194 mb
dCas9-TET:gRNA3 v.
dCas9:gRNA3:Cre
dCas9-TETgRNAscr v.
dCas9:gRNAscr:Cre
dCas9:gRNA3:Cre v.
dCas9:gRNAscr:Cre
Mouse genes
e
-log10(q)
Il33-002 CpG 11
q = 3.03 x 10-6
Il33-002 CpG 10
q = 2.53 x 10-5
chr8:4802686
q = 1.19 x 10-4
q = 0.01
Chromosome
f
gRNA3 CGACTGTGTTAGCTCTCCACNGG
chr8:48026756-48026856 GTTGTCAGTCTGATGATCAGGAG
gRNA3 CGACTGTGTTAGCTCTCCACNGG
chr8:125401335-125401435 GCGTTACCCAAGCCCTCCACAGG
gRNA3 CGACTGTGTTAGCTCTCCACNGG
chr17:66124338-66124438 CTGCTAGGCGAGGGCTCCACTAG
gRNA3 CGACTGTGTTAGCTCTCCACNGG
chr3:105661468-105661568 TGAATGTGTCACATCTACCCTGG
g
-log10(FDR)
Legend
FDR >0.05
FDR<=0.05
log Fold Change [log2(gRNAscr) - log2(gRNA3)]
Enriched in gRNA3
Enriched in gRNAscr
h
IL33-002
Peak
Peak summit
gRNA3
i
FDR Rank: 1
chr11:54,223,607-54,224,007
Seq: GGATCTCCTCTCCTAGCTCTCCACAGG
gRNA3: GGA----CTGTGTTAGCTCTCCACNGG
FDR Rank: 2
chr4:8,906,651-8,907,051
Seq: TGACTGCAACTAGCTCTCCACAGG
gRNA3: GGACTG-TGTTAGCTCTCCACNGG
FDR Rank: 3
chr16:10,452,580-10,452,980
Seq: TGAGCCACACCAGCTCTCCACGGG
gRNA3: CGA-CTGTGTTAGCTCTCCACNGG
FDR Rank: 4
chr8:114,673,330-114,673,730
Seq: AGAAGACAGCAGCTCTCCACTGG
gRNA3: GGACTGTGTTAGCTCTCCACNGG
FDR Rank: 5
chr2:165,747,915-165,748,315
Seq: GGCAGAGCTACAAGCTCTCCACCGG
gRNA3: GG-A-CTGTGTTAGCTCTCCACNGG
j
-log10(FDR)
% Methylation Change (gRNA3 - gRNAscr)

Figure 6. WGBS and ChIP-seq analyses of dCas9 and dCas9-TET approaches to targeted demethylation. (A) Clustering of NIH-3T3 samples with indicated lentiviral treatment and replicate number by CpG methylation, based on highly covered (≥10X) CpGs common to all samples with the cluster using Samples function in the methylKit package for R (ward.D2 method). (B) Fraction of total sequenced CpGs (mean ± SEM) that read as methylated (C after bisulfite conversion) in each treatment type, aligned and calculated with Bismark default parameters (*n* = 3, biological replicates; ** indicates *P* < 0.01 with two-sided Student's t-test). (C) Number of significantly differentially methylated CpGs (dmCpGs) (red = hypomethylated, blue = hypermethylated) determined by methylKit calculateDiffMeth function (≥5X coverage, *n* = 3, q-value (p-value adjusted for multiple testing by SLIM method) < 0.01, 25% methylation difference) of dCas9TET:gRNAscr and dCas9TET:gRNA3 NIH-3T3 cell lines compared to untreated control NIH-3T3 cells (left) or compared to dCas9:gRNAscr:Cre or dCas9:gRNA3:Cre, respectively. (D) Genome browser view of mouse (mm10 genome) chromosome 1, with bedGraphs containing hypomethylated dmCpGs (and amount of hypomethylation in %) in dCas9-TET:gRNA3 (top, blue) and dCas9-TET:gRNAscr (middle, light blue) from (C, right, chromosome 1 only) and dCas9:gRNA3:Cre hypomethylated dmCpGs compared to dCas9:gRNAscr (bottom, pink), which are the same as pink inset in (E). Range is 0 to -100. Gene structures are densely mapped and sparsely labeled at the bottom (dark blue) (E) Manhattan plot of all hypomethylated (>25% change in methylation) sites in dCas9:gRNA3:Cre cells compared to dCas9:gRNAscr:Cre. Significantly differentially methylated sites were considered under default methylKit conditions (*q* < 0.01, above horizontal blue line). CpGs circled in red and labeled with q-values represent two target *Il33-002* CpGs (10 and 11) and the CpG circled in black represents the third-highest (top non-target) dmCpG ranked by q-value. Pink box highlights significant dmCpGs in chromosome 1, which are displayed in (D). (F) Best alignments to gRNA3 and PAM sequence of four 100bp regions surrounding (+/-50bp) four selected off-target dmCpGs (of 641 total) from (E). Dashes indicate mismatches. Vertical lines indicate matching base pairs. Matched base pairs in off-

target region are also shown in blue. For the purposes of this representation, the adenine in the NAG PAM in considered a mismatch to the more active guanine in the NGG PAM. Top alignment (in black) shows the region containing the top off-target dmCpG by q-value, chr8:4802686, circled in black in (E) (14/23 mismatches). Second from top alignment displays the 100bp off-target region containing a sequence with the most similarity to gRNA3 of all 641 100bp off-target regions, as calculated by position-weighted mismatch algorithm CCTop (10 mismatches). Third from top alignment displays off-target region ranked second for similarity to gRNA3 by CCTop (8 mismatches but one is closer to 3'/PAM end than above). Final alignment shows the off-target region with the lowest mismatches overall to gRNA3, regardless of position, which is 7. (G) Volcano plot of ChIP-seq significantly differentially enriched regions in gRNA3 and gRNAscr conditions (*n* = 3) for anti-FLAG ChIP-seq against FLAG-dCas9, using input as control. Log2 fold change (log2(gRNAscr) - log2(gRNA3)) is plotted on the x-axis and -log10(False Discovery Rate) is plotted on the y-axis. The locus corresponding to the *Il33* transcription start site is circled in red. (H) Genome browser view (mm10) of *Il33-002* (blue). Statistically significant peak (turquoise) (circled in red in (G)), peak summit (purple) and gRNA3 sequence (pink) are labeled. (I) Manually curated sequence alignments of top 5 DERs from ChIP-seq data to gRNA3 and PAM sequence. Identical sequence matches are marked in blue and bolded. Potential gaps are indicated with dashes. Chromosomal locations are given for the mm10 genome build. (J) Volcano plot depicting changes in methylation and associated statistical probabilities in dCas9:gRNA3 NIH-3T3 cells (from WGBS data) for CpGs that are within 150 differentially enriched off-target regions bound by dCas9:gRNA3 (*Il33* is excluded). Change in methylation (x-axis) is expressed as mean percent methylation in dCas9:gRNAscr subtracted from mean percent methylation in dCas9:gRNA3. Statistical probabilities are provided as the -log10 of the p-value derived by the independent t-test as corrected for multiple testing by the False Discovery Rate method. 549 CpGs were located in DERs after filtering for CpGs that were 5X covered in all 6 samples; however, 132 CpGs with exactly 0% methylation in all 6 samples and 13 CpGs with exactly 100%

methylation in all 6 samples are not depicted as p-values cannot be mathematically calculated. Therefore, 404 CpGs are shown. Underlying WGBS data was presented above and reflects *n* = 3 independent stable cell lines.

For all significantly differentially hypomethylated CpGs (differentially methylated CpGs, dmCpGs), we defined the following thresholds: covered ≥5X in all replicates from treatments in question, q-value (SLIM-adjusted p-value) less than 0.01, and a difference in methylation ≥25%. When comparing all treatment conditions to untreated controls, there were 54 dmCpGs in dCas9:gRNAscr, 338 in dCas9:gRNA3, 3,940 in dCas9-TET:gRNAscr, and 6,286 in dCas9-TET:gRNA3. Due to differences in sample-level read depth (**Supplementary Table 4**), the direct comparison of the numbers of dmCpGs could suffer from coverage bias, therefore, the incidence of dmCpGs as a fraction of all CpGs assayed (≥5X covered in all 6 samples) under each comparison to untreated cells are as follows: dCas9:gRNAscr = 54/9,039,707 (0.0006%), dCas9:gRNA3 = 338/9,903,308 (0.0034%), dCas9-TET:gRNAscr = 3,940/7,503,634 (0.0525%), dCas9-TET:gRNA3 = 6286/10,931,608 (0.0575%). Accordingly, after this normalization, dCas9-TET produces 16.9X (gRNA3) to 87.5X (gRNAscr) more demethylated CpGs. Furthermore, as dCas9:gRNA cells serve as better experimental controls (e.g. for lentiviral integration) than untreated controls, comparisons of dCas9-TET cells to dCas9 cells expressing the same gRNA are more appropriate and result in the following numbers of hypomethylated dmCpGs: dCas9-TET:gRNAscr = 26,860/8,216,634 (0.3269%), dCas9-TET:gRNA3 = 98,568/13,290,423 (0.7416%) (**Figure 6C** right, **D** top and middle panel). These data emphasize the genomic hypomethylation burden of dCas9-TET and establish that genomic hypomethylation of the dCas9 demethylation method to be far more limited.

Next, we defined off-targets of dCas9:gRNA3 by comparison to dCas9:scr under the same minimum coverage and statistical conditions (**Figure 6E,D** bottom panel) and found a total of 643 dmCpGs (**Supplementary Data 1**). Interestingly, the top 2 dmCpGs in terms of statistical significance were the target *Il33-002* CpGs 10 ($q=2.53 \times 10^{-5}$) and 11 ($q=3.03 \times 10^{-6}$) (**Figure 6E, circled in red**). Upon further inspection, the third target CpG, CpG 9, failed to be identified in this analysis because it was 4X covered in one sample (**Supplementary Table 5**) but was significantly demethylated in this dataset

(87.45% v. 8.15% *P* = 0.0011, t-test). The highest ranked off-target dmCpG was chr8:4802686 (**Figure 6E**, circled in black), yet the highest scoring sequence match to the gRNA3 target within +/- 50bp from this CpG (using CCTop [387], an algorithm that identifies and ranks off-targets on both DNA strands by position and number of mismatches) had 13 mismatches to gRNA3 (**Figure 6F**, black panel), including one in the most deleterious position, immediately adjacent to the PAM, and a non-standard NAG PAM, making it unlikely to be demethylated as a consequence of off-target dCas9 binding.

To see if any of the 641 off-target dmCpGs had any sequence similarity to gRNA3, we first compiled a comprehensive list of 100bp regions surrounding all possible gRNA3 off-targets in the murine genome of up to 4 mismatches and 1 gap (4,436 total), representing what is typically accepted to be the maximum number of tolerated mismatches by CRISPR/Cas9 [387] (generated by combining lists from 4 online tools CRISPR DESIGN (crispr.mit.edu; deprecated), OFF-Spotter [388], CCTop [387], and OFF-Finder[389]) and searched this list for the presence of the dCas9 off-target dmCpGs. None of the 4,436 potential off-target sites overlapped with equal-sized 100bp regions containing the significantly hypomethylated off-target CpGs. In an effort to find any similarity of dmCpGs to gRNA3, we again invoked CCTop to search the list of 100bp regions surrounding the dmCpGs to identify the highest sequence similarity to gRNA3 and found the following top off-candidate regions (**Figure 6F**): chr8:125401335-125401435 with 10 mismatches but a complete 6-bp match to the 3' (seed) region and 9 of 10 matches to the 10 most 3' nucleotides and chr17:66124338-66124438 with a similar 6bp complete 3' match, only 8 mismatches, but an NAG pam instead of NGG PAM. The fewest possible mismatches to any sequence within the 100bp dmCpG-containing regions was 7, but this scored lower as it included 2 mismatches 2 and 4bp from the PAM, which is not compatible with dCas9 binding. The complete list of CCTop-generated mismatched (up to 18 mismatches) off-targets is available in **Supplementary Data 2**.

Given the facts that dCas9 is much less tolerant to mismatches in the seed region (5-12 bp nearest to the PAM), NAG PAMs display an estimated one-fifth to one-tenth of NGG PAM activity, and most importantly, that although there is more tolerance for mismatches in the 5' region there seem to be no reports in the literature of activity with more than 5 mismatches anywhere in the sequence [390,391], we hypothesized that none of the dmCpGs are genuine off-targets of gRNA3. To test this hypothesis, we performed chromatin immunoprecipitation sequencing (ChIP-seq) with an anti-FLAG antibody in cells expressing FLAG-tagged dCas9 and either gRNA3 or gRNAscr. We found 151 significantly differentially enriched regions (DERs) of dCas9:gRNA3 (FDR <0.05) and 44 DERs in dCas9:gRNAscr (**Figure 6G**, **Supplementary Data 3**). The most enriched locus (by fold change) was the targeted *Il33* promoter and the summit of the peak (highest fragment pileup and predicted binding spot) was within the gRNA3 target sequence (**Figure 6H**). Other DERs included 6 of the 4,436 off-target sites predicted above. Manual analysis of the top 5 gRNA3 DERs (sorted by FDR) revealed considerable sequence similarity to gRNA3, with 100% alignment of a 10-11 bp seed region and PAM (**Figure 6I**). Importantly, none of the 150 (excluding the *Il33-002* DER) gRNA3 DERs overlapped with the 641 DMGs from the WGBS data: the only region demethylated and bound by dCas9 was *Il33-002*, reinforcing the high specificity of this approach. Moreover, restricting differential methylation analysis to only the 549 CpGs (with a minimum coverage of 5X in all 6 samples) located within the 150 off-target DERs bound by dCas9:gRNA3 revealed no statistically significant differentially methylated sites and no otherwise apparent trend that favors nonsignificant hypomethylation over hypermethylation (**Figure 6J**).

These data suggest that the dmCpGs may originate from an activity that is not the off-target binding of dCas9:gRNA, such as differential epigenetic drift during cell passaging[392], global epigenetic change as a response to lentiviral integration[385,386], technical variability in WGBS, or by insertional mutagenesis and lentiviral integration

into gene-regulatory elements that could also lead to modified expression of epigenetic editing enzymes[393]. To address some of these potential factors, we analyzed split sequencing reads from our WGBS data (see Methods and Supplementary Software 1) to identify 2,792 possible lentiviral insertion points across all 6 replicates and found that of 641 off-target dmCpGs, 13 are within +/- 5kb from viral insertion sites (P = 0.0322, hypergeometric) and 97 are within +/- 50kb (P = 0.00729, hypergeometric). Additionally, as it is known that lentivirus integration predominantly results in genomic hypermethylation[385,386], we wondered if dCas9 off-target dmCpGs were identified as hypomethylated in dCas9:gRNA3:Cre because these sites were aberrantly hypermethylated in dCas9:gRNAscr:Cre, rather than by direct demethylation in dCas:gRNA3:Cre cells. Indeed, 3 dmCpGs, including the top hypomethylated by q-value – chr8:4802686 (**Figure 6E**, circled in black) – were significantly hypermethylated dmCpGs in dCas9:gRNAscr:Cre as compared to untreated control cells. As a lack of statistical significance in these sites compared to untreated does not discount a lack of statistical significance compared to dCas9:gRNA3:Cre (for example, these sites can show less variability in dCas9:gRNA3:Cre than untreated), we were prompted to see what fraction of the 641 off-target dmCpGs were generally hypermethylated in dCas9:gRNAscr:Cre as compared to untreated control cells. Of the 641 dmCpGs, 424 were sufficiently covered (≥5X) in all six dCas9:gRNAscr:Cre and untreated samples. Of these 424 sites, 379 (89%) were generally hypermethylated. 246 of these (65%) were nominally significant (P < .05, one-sided t-test) and 179 were still significant after correction for multiple testing (false discovery rate) (**Supplementary Data 4**).

We also used targeted-bisulfite pyrosequencing to assess whether dCas9:gRNA3 caused demethylation of the top 5 predicted candidate off-target CpGs for gRNA3 and found that there was no observable change in methylation of any of the top-predicted off-targets (**Supplementary Figure 3C**, **Supplementary Table 6**).
Interestingly, under the same analysis conditions, there were no significantly differentially methylated CpGs between dCas9-TET:gRNA3 and dCas9-TET:gRNAscr,

further emphasizing the non-specific activity of the TET domain even when it is targeted by the CRISPR system. To provide further evidence that the dCas9-TET hypomethylated dmCpGs (**Figure 6C**, right) might originate as a consequence of TET-directed (rather than dCas9-directed) interaction with DNA of the dCas9-TET fusion protein, we analyzed whether these dmCpGs are enriched in established sites of TET action: enhancers [394-399]. In dCas9TET:gRNA3 cells, 815 of 106,966 dmCpGs (0.76%) could be found in mouse enhancers (FANTOM5 project [400], mouse_permissive_enhancers_phase_1_and_2.bed.gz) compared to 89,922 of all 13,290,423 ≥5X covered CpGs (0.68%) (P = 2.93 x 10^{-5}, hypergeometric). There was also a significant enrichment of dmCpGs in enhancers in dCas9TET:gRNAscr cells, where 244 of 26,860 of dmCpGs were in enhancers while 56,417 of all 8,216,634 ≥5X covered CpGs (0.91% v. 0.69%) were in enhancers (p = 3.03 x 10^{-6}, hypergeometric). Importantly, an even greater fraction (46 of 4174 or 1.1%) of shared dmCpGs between dCas9TET:gRNA3 and dCas9TET:gRNAscr were found in enhancers. Of all regions containing predicted gRNA3 off-targets of up to 4 mismatches and 1 gap (100bp around cut site), 21 and 5 were within 100bp of dmCpGs in dCas9-TET:gRNA3 and dCas9-TET:gRNAscr, respectively. 45 and 24 of these gRNA3 and gRNAscr dmCpGs, respectively, were bound by dCas9 in the ChIP-seq data.

dCas9-based demethylation analysis of the role of TSS methylation in *SERPINB5*, *Tnf* and *FMR1* genes

Our previous results show that methylation of *Il33-002* TSS silences basal promoter activity but that demethylation does not result in robust activation of the gene. Induction of this gene could occur independently of methylation of the promoter. We therefore examined whether TSS (de)methylation might play similar or different roles in other genes.

We next examined the *SERPINB5* gene, which encodes the tumor suppressor maspin and is methylated and transcriptionally silenced in human MDA-MB-231 breast cancer cells. Reactivation of this gene has been reported to increase cell adhesion and therefore decrease growth, invasion, and angiogenesis [401-405]. Several studies have reported that DNA methylation of the *SERPINB5* promoter negatively correlated with gene expression in human cancer and that 5-aza-2'-deoxycytidine treatment is sufficient to restore *SERPINB5* expression [406-410].

We designed a single gRNA targeting 6 CpGs (3 within the gRNA binding site and 3 within 11bp of the 3' end of the gRNA, as predicted to be completely affected by our *in vitro* footprint assays in **Figure 3**) in the core *SERPINB5* promoter and specifically in the transcription-regulatory GC-box (**Figure 7A**). In this case, increasing puromycin had a mild effect in increasing the frequency of unmethylated promoters and even the highest puromycin concentrations (40 µg/mL) resulted in demethylation of only 20% (**Supplementary Figure 8**). We reasoned that perhaps there is a strong selection against cells expressing *SERPINB5* – which is a known tumor suppressor – resulting in overgrowth of cells bearing highly methylated *SERPINB5*. Therefore, we turned to the previously described clonal isolation strategy. We picked approximately 20 clones from each of the two treatments (gRNAscr and gRNASERPINB5) and evaluated methylation by pyrosequencing, which revealed a significant demethylation in gRNASERPINB5 MDA-MB-231 clones on average comparted to gRNAscr clones (**Figure 7B**). We found that numerous clones were completely demethylated (**Figure 7C**) and we selected 5 gRNASERPINB5 clones with methylation levels below 5% at all six CpGs as well as 5 representative gRNAscr clones.

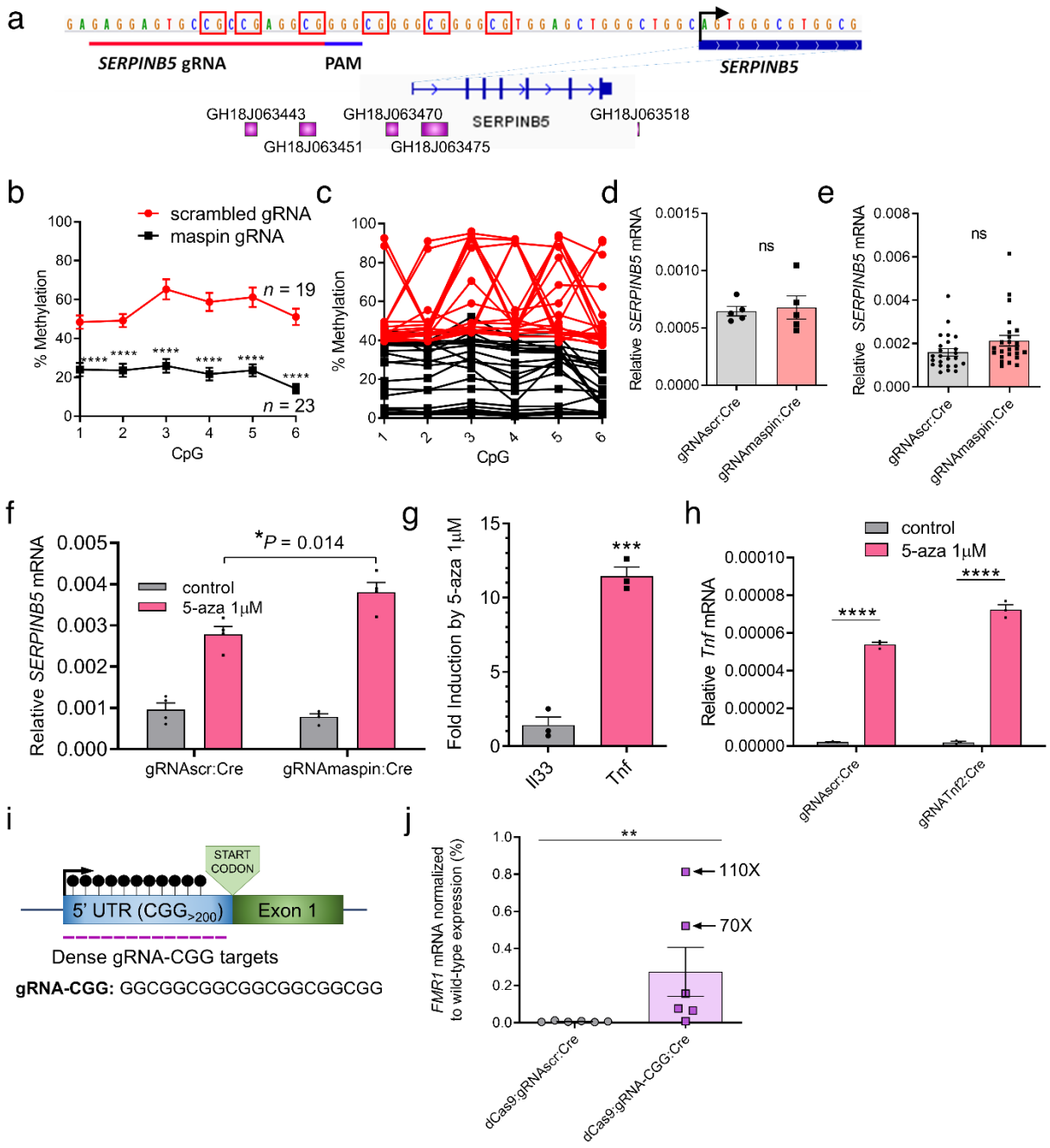

a
SERPINB5 gRNA
PAM
SERPINB5
GH18J063443
GH18J063470
SERPINB5
GH18J063518
GH18J063451
GH18J063475
b
scrambled gRNA
maspin gRNA
% Methylation
n = 19
n = 23
CpG
c
% Methylation
CpG
d
Relative SERPINB5 mRNA
ns
gRNAscr:Cre
gRNAmaspin:Cre
e
Relative SERPINB5 mRNA
ns
gRNAscr:Cre
gRNAmaspin:Cre
f
Relative SERPINB5 mRNA
*P = 0.014
control
5-aza 1μM
gRNAscr:Cre
gRNAmaspin:Cre
g
Fold Induction by 5-aza 1μM
Il33
Tnf
h
control
5-aza 1μM
Relative Tnf mRNA
gRNAscr:Cre
gRNATnf2:Cre
i
START CODON
5' UTR (CGG>200)
Exon 1
Dense gRNA-CGG targets
gRNA-CGG: GGCGGCGGCGGCGGCGGCGG
j
FMR1 mRNA normalized to wild-type expression (%)
110X
70X
dCas9:gRNAscr:Cre
dCas9:gRNA-CGG:Cre

Figure 7. **The effect of dCas9-based demethylation of TSS on expression of *SerpinB5*, *Tnf* and *FMR1* genes.** (A) (Top) Schematic of the human *SERPINB5* promoter region, including the start site of transcription (marked by black arrow) and the binding site and PAM of the *SERPINB5* gRNA. CG sequences are boxed in red. (Bottom) *SERPINB5* gene with purple boxes indicating enhancer positions relative to gene body. Enhancer IDs correspond to the GeneHancer database. (B) DNA methylation level of each CpG averaged over *n* = 19 gRNAscr (red) and *n* = 23 gRNASERPINB5 (black) independent MDA-MB-231 clones isolated from 3 independent treatments of cell cultures as assessed by pyrosequencing (mean ± SEM). (C) Same data as (B) except now shown as the calculated methylation fraction for each of the 19 gRNAscr (red) and 23 gRNASERPINB5 (black) clones, rather than the average of all clones. (D) *SERPINB5* expression levels measured by RT-qPCR and normalized to *GAPDH* expression levels for 5 gRNAscr and 5 lowly-methylated gRNASERPINB5 clones (mean ± SEM, *n* = 5 biologically independent clones). (E) *SERPINB5* expression levels (mean ± SEM) measured by RT-qPCR and normalized to *GAPDH* expression levels for 48 (*n* = 24 for each treatment) MDA-MB-231 clones subcloned from the clones in (D). (F) *SERPINB5* expression levels (mean ± SEM) measured by RT-qPCR and normalized to *GAPDH* expression levels for clones from (D) following treatment with 1 μM 5-aza-2'-deoxycytidine or water control (*n* = 5 biologically independent experiments). (G) Expression fold change of murine *Il33-002* (grey) and *Tnf* (pink), normalized to *Actb* and water control (mean ± SEM), following treatment of control NIH-3T3 cells with 1 μM 5-aza-2'-deoxycytidine (*n* = 3 biologically independent experiments). (H) *Tnf* expression (mean ± SEM) in NIH-3T3 cell lines stably in control (water); grey bars) or 1 μM 5-aza-2'-deoxycytidine (pink bars) expressing either gRNAscr or gRNATnf2:Cre under high-puromycin conditions in combination with dCas9, followed by dCas9 removal by Cre recombinase, as assayed by RT-qPCR and normalized to *Actb* expression (*n* = 3 biologically independent experiments). (I) Schematic of the human *FMR1* repeat region showing the 5' untranslated region (UTR) that is prone to CGG repeat expansion and methylation in Fragile X syndrome.

Sequence of the gRNA targeting this region is shown (gRNA-CGG) and the extent of the available binding sites for this gRNA is represented by purple lines which indicate binding sites, the 13 presented here represent less than 15% of the available binding site in the Fragile X syndrome patient primary fibroblasts used in this study, which have approximately 700 CGG repeats. (J) *FMR1* expression quantified by RT-qPCR and normalized to *GAPDH* expression levels in Fragile X syndrome patient primary fibroblasts that had stably expressed dCas9 (later removed with Cre) and either gRNAscr (grey) or gRNA-CGG (purple) under high-puromycin selection (n = 6 biologically independent experiments, mean ± SEM). Data is represented as a percent of the expression of *FMR1* in wild-type age-matched primary fibroblasts (Mann-Whitney test, two-sided). * indicates statistically significant difference of $P < 0.05$, ** of $P < 0.01$, *** of $P < 0.001$, **** of $P < 0.0001$, and ns = not significant (Student's t-test, two-sided, with Holm-Sidak correction if number of tests is greater than 3). Exceptionally, for (J) Mann-Whitney test was used due to unequal variance.

Methylation levels of 3 gRNAscr clones and 3 gRNASERPINB5 clones remained constant for at least 45 additional days of passaging, though there appeared to be a small non-significant trend of increasing methylation in demethylated gRNASERPINB5 clones (**Supplementary Figure 8B-C**). Surprisingly, despite the large change in methylation, *SERPINB5* expression after Cre-mediated dCas9 removal remained unchanged between the two sets of clones, though there was a small insignificant (P = 0.105) increase in the variance of expression levels in the different demethylated clones (**Figure 7D**). The difference in *SERPINB5* expression was increased when these cells were further subcloned in order to reduce the potential of selection against cells with activated *SERPINB5* expression (**Figure 7E**), but not to a statistically significant degree (P = 0.0767), suggesting that demethylation of the *SERPINB5* promoter is insufficient to activate the gene. Since 5-aza-2'-deoxycytidine was shown to induce the gene we tested whether induction of the gene requires additional demethylation beyond the gene TSS: we tested whether 5-aza-2'-deoxycytidine would induce the methylated and unmethylated *SERPINB5* promoter to the same extent. In contrast to *Il33-002*, which was not further induced by 5-aza-2'-deoxycytidine after TSS demethylation, expression of *SERPINB5* with a demethylated TSS region was significantly increased by 5-aza-2'-deoxycytidine treatment as compared to gRNAscr cells treated with 5-aza-2'-deoxycytidine (**Figure 7F**) (P = 0.0184) (4.85X in gRNASERPINB5 vs. 2.59X in gRNAscr). This is consistent with the conclusion that demethylation of the promoter is insufficient for its expression and demethylation of other regions, such as the depicted enhancer regions (**Figure 7A**), is required for induction of *SERPINB5*; however, basal promoter demethylation contributes to the overall expression level following demethylation of other regions.

We then questioned whether larger changes in expression could follow demethylation of proximal promoters in other genes. To identify genes that may potentially display such changes, we selected 17 candidate genes in NIH-3T3 cells with large expression fold changes in response to 5-aza-2'-deoxycytidine in a publicly available microarray dataset

(GEO GSE8374) and analyzed their expression changes by RT-qPCR following 1μM 5-aza-2'-deoxycytidine treatment (**Supplementary Table 7**). We selected the *Tnf* gene which was heavily methylated at the proximal promoter region and the expression of which was increased by more than ten-fold by 5-aza-2'-deoxycytidine treatment (**Figure 7H**). We tested six gRNAs under high-puromycin selection (20 μg/mL) conditions and identified a gRNA that demethylated all 10 CpGs in approximately 200bp upstream of the *Tnf* TSS (**Supplementary Figure 9A-C**). We chose this gRNA (gRNATnf2) for Cre recombinase removal of dCas9. Surprisingly, complete *Tnf* promoter demethylation did not result in a significant difference in *Tnf* expression compared to gRNAscr (**Figure 7H**) nor could we observe any difference in expression in subclones from these cell pools (**Supplementary Figure 8D**). However, when these cells were treated with 5-aza-2'-deoxycytidine, the demethylated gRNATnf2 cells were induced to a larger extent than the methylated gRNAscr pools (36-fold versus 24-fold) (P = 0.0008) (Figure 7H). Therefore, we conclude that, similar to demethylation of *SERPINB5* TSS, demethylation of *Tnf* basal promoter contributes to expression but is insufficient to induce expression and that expression necessitates demethylation of a different region either in *cis*, such as the two murine proximal *Tnf* enhancers (**Supplementary Figure 9A**)[411], or in *trans* through activation of putative transcription factors.

Our final demethylation target was the *FMR1* gene which, in patients with Fragile X syndrome, undergoes a CGG repeat expansion (>200 repeats) in its 5' UTR that becomes aberrantly hypermethylated and results in silencing of *FMR1* transcription[412]. This region has repeatedly been shown to be reactivated by 5-aza-2'-deoxycytidine[413-416] and we validated it herein (**Supplementary Figure 10A**). The CGG repeat expansion is a unique target for a guide RNA with the sequence GGCGGCGGCGGCGGCGGCGG and PAM motif CGG since it should bind sequentially to the entire large repeat region and – under sufficient expression levels – shield the entire region from methyltransferase activity (**Figure 7I**). We obtained publicly available primary fibroblasts from a patient with Fragile X syndrome with approximately 700 CGG

repeats exhibiting high methylation, as determined previously [417], – and a lentiviral vector bearing the CGG-targeting gRNA sequence (gRNA-CGG) [418]. After application of our optimized dCas9-demethylation protocol using gRNA-CGG or gRNAscr (20 µg/mL puromycin) we observed a reduction in the methylated CGG repeat fraction in the gRNA-CGG condition (Supplementary **Figure 10B-D**) and significant upregulation of *FMR1* gene expression (P = 0.0087) **(Figure 7J**), characterized by an increase from a mean 0.7% of wild-type expression in gRNAscr cells to a mean of 27% in gRNA-CGG cells and as much as a 110-fold induction in one cell line corresponding to 81% of wild-type *FMR1* levels. The magnitudes of induction of *FMR1* gene expression are vastly larger than the induction following TSS demethylation observed in *Il33* and are suggestive of the fact that in this case DNA methylation of the repeat region has a large effect on gene expression.

In summary, we demonstrate that the dCas9 demethylation method can be effectively applied in several different cell types: a murine fibroblast cell line, a human breast cancer cell line, and primary patient fibroblasts and across different genetic contexts. This method could be used to assess the relative contribution of DNA methylation in specific sites to modulation of gene expression and to delineate positions whose demethylation would have the largest effect on expression. Since our method physically targets DNA methylation without confounding enzymatic activities it provides an unconfounded and at times surprising assessment of the role of DNA methylation.

CRISPR/Cas9-induced demethylation confounds mutational studies with Cas9

The catalytically active CRISPR/Cas9 system has become the gold standard technique for generating gene knockouts in functional studies. A common technical consideration in these approaches is to target 5' constitutive exons such that frameshift mutations are more likely to take effect early and render the translated protein non-functional[306]. This inevitably results in the positioning of the Cas9:gRNA ribonucleoprotein complex near

the TSS and proximal promoter of the targeted gene. Based on the results describe here, we hypothesized that the residence time of DNMT-interfering Cas9, in addition to the drastic epigenetic changes that occur during post-mutagenesis repair[419], may in certain cell subpopulations result in DNA demethylation and gene induction that would confound the interpretation of the results.

We had in a previous study used Cas9 and an *HNF4A*-targeting gRNA from the commonly used GECKO gRNA library[306] to generate *HNF4A* gene knockouts in primary human hepatocytes [291]. The gRNA target site is located in the first exon of several *HNF4A* isoforms, the *HNF4A* TSS is only 2 bp from the 3' end of the PAM, and there are 3 CpGs directly within the site, with two additional CpGs in close proximity (**Figure 8A**). We analyzed one mixed *HNF4A* CRISPR:Cas9 targeted cell population and mapped by Sanger sequencing different *HNF4A* alleles, which were primarily bearing a T->C missense mutation as well as in-frame and out-of-frame deletions (**Figure 8A-C**), indicating that a considerable fraction of cells in this population were likely to produce a protein that retained some degree of functionality. To our surprise, we found that this highly methylated region was completely demethylated in this cell population, irrespective of the mutation induced by Cas9 (**Figure 8D**). This demethylation was both substantial and broad, covering not just a 311bp fragment with 15 CpGs highly methylated in gRNAscr cells to over 90% on average, but also continued to a slightly smaller degree into originally less methylated regions immediately upstream (230bp with 5 CpGs) and downstream (269bp with 7 CpGs) (**Figure 8E**). We also found that this demethylated gHNF4A population expressed approximately 15-fold more *HNF4A* mRNA than gRNAscr controls (**Figure 8F**). Thus, standard CRISPR/Cas9 gene depletion studies might be confounded by the effects of extensive demethylation.

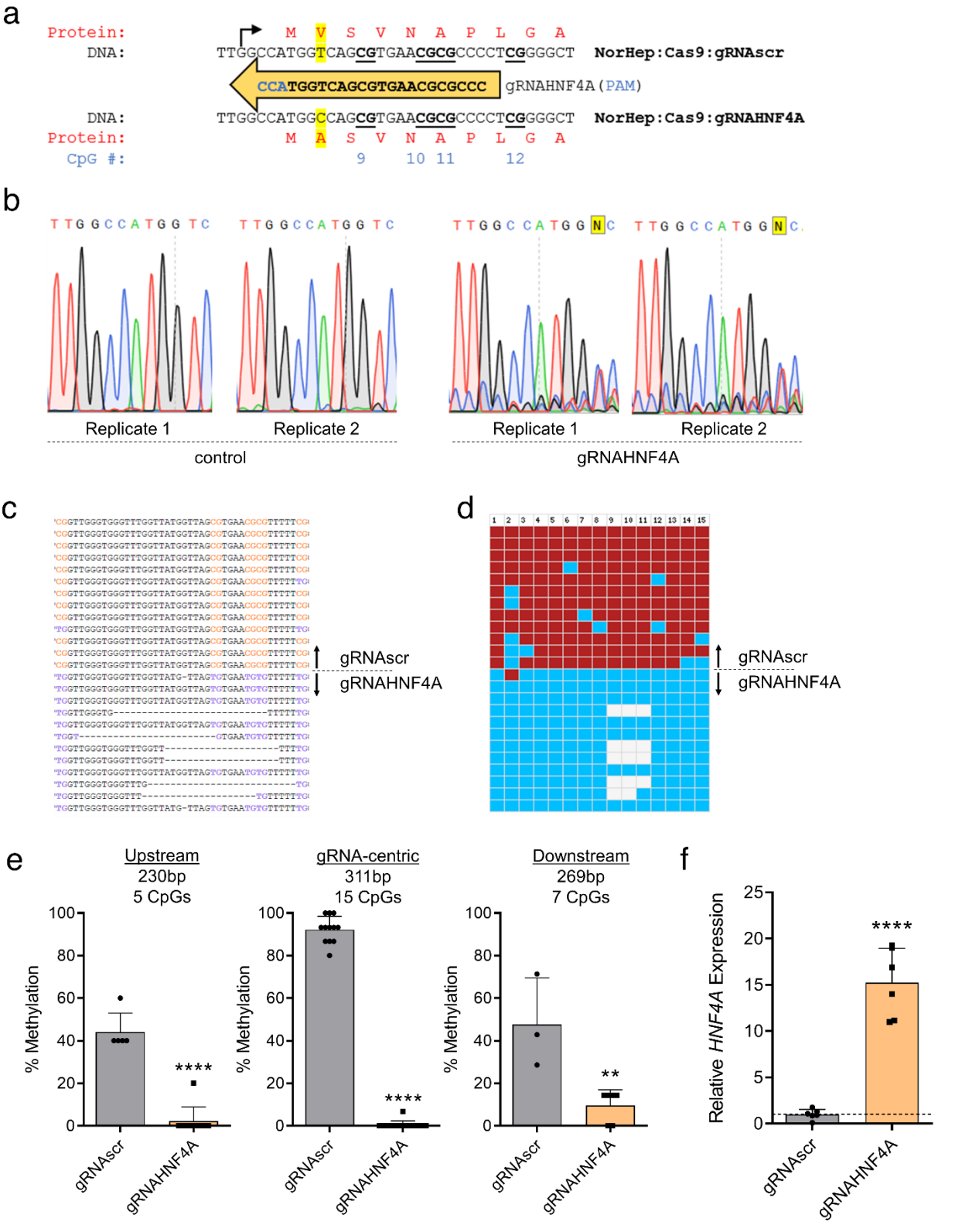

a
Protein: M V S V N A P L G A
DNA: TTGGCCATGGTCAGCGTGAACGCGCCCCTCGGGGCT NorHep:Cas9:gRNAscr
CCATGGTCAGCGTGAACGCGCCC gRNAHNF4A(PAM)
DNA: TTGGCCATGGCCAGCGTGAACGCGCCCCTCGGGGCT NorHep:Cas9:gRNAHNF4A
Protein: M A S V N A P L G A
CpG #: 9 10 11 12
b
Replicate 1
Replicate 2
control
Replicate 1
Replicate 2
gRNAHNF4A
c
gRNAscr
gRNAHNF4A
d
gRNAscr
gRNAHNF4A
e
Upstream
230bp
5 CpGs
gRNA-centric
311bp
15 CpGs
Downstream
269bp
7 CpGs
% Methylation
gRNAscr
gRNAHNF4A
****
****
**
f
Relative HNF4A Expression
gRNAscr
gRNAHNF4A
****

Figure 8. Demethylation is a confound of Cas9 knockout gene deletion. (A) The sequence of the lentiviral gRNAHNF4A and its PAM site in blue. Above is the reference sequence of the *HNF4A* gene near the gRNA target site, as validated by Sanger sequencing in primary human hepatocytes expressing via lentivirus Cas9 and gRNAscr, with the TSS indicated by a black arrow and the reference protein sequence in red. CGs are bolded and underlined. Below is the dominant Sanger sequence profile of a primary human hepatocyte population expressing lentiviral Cas9 and gRNAHNF4A. This mutation and the resulting difference in the amino acid sequence, as well as the reference sequences at this location, are highlighted in yellow. (B) Two technical replicates each of the Sanger sequencing chromatograms from the primary human hepatocytes expressing dCas9 and gRNAscr (left) or dCas9 and gRNAHNF4A (right) at the targeted *HNF4A* locus. (C) Sanger sequencing results of 13 gRNAscr and 12 gRNAHNF4A DNA strands following bisulfite conversion from the cell populations in (B), demonstrating both the methylation levels and the variety of mutations induced by Cas9 in gHNF4A-treated cells. (D) Same as (C) except data expanded is expanded to a larger (>300bp) region, and simplified such that only CpGs are shown, where blue squares indicate unmethylated CpGs, red squares indicate methylated CpGs, and white squares indicate missing information due to Cas9-induced deletions. CpGs are numbered in accordance with (A). (E) Bisulfite-sequencing data from (D) (center) as well as 5 CpGs immediately upstream (left) and 7 CpGs immediately downstream (right), displayed as percent DNA methylation over all sequenced DNA strands in primary human hepatocytes expressing Cas9 and either gRNAscr (grey) or gRNAHNF4A (orange) and as mean ± SD as it is summary data from one mutated cell line. Individual dots represent individual strands of DNA from this clonal cell line. (F) *HNF4A* expression in primary human hepatocytes expressing Cas9 and either gRNAscr (grey) or gRNAHNF4A (orange) quantified by RT-qPCR and normalized to *GAPDH* expression, followed by normalization to average expression in gRNAscr cells, with a dashed line at 1 (n = 6 independent clones, mean ± SD). * indicates statistically significant difference of $P < 0.05$, ** of $P < 0.01$, *** of $P < 0.001$, **** of $P < 0.0001$,

and ns = not significant (Student’s t-test, two-sided, with Holm-Sidak correction if number of tests is greater than 3).

12.4 Discussion

The developmental profiles of DNA methylation across human tissues[420] combined with the fact that deviations from these patterns are associated with disease[350,421,422] suggest that DNA methylation has an important role in physiological processes. Importantly, it has been suggested that DNA methylation plays a functional role in the molecular pathology of cancer[350,352,408,422-424] and other common diseases, including mental health disorders[425-427].

Correlation studies since the early 1980s have suggested that DNA methylation in promoters and other transcriptional regulatory regions is negatively correlated with gene expression[428-431]. In the last three decades, several lines of evidence have provided support to the causal role of DNA methylation in the modulation of gene expression. First, *in vitro* methylation of reporter plasmids was shown to silence transcriptional activity when these plasmids were transfected into cell lines[430]. Later studies used different methods to limit *in vitro* methylation to specific regions. Although these studies provide the most direct evidence that there are cellular mechanisms to recognize DNA methylation in particular regions and translate this into silencing of gene activity, the main limitation of these studies is that silencing of ectopically methylated DNA might not reflect on genomic methylated sites and might instead represent a defense mechanism to silence invading viral and retroviral DNA[432] rather than a mechanism for cell-type-specific differential gene expression. Second, DNA methylation inhibitors 5-aza-2'-deoxycytidine and 5-azacytidine provided early evidence for a causal role for DNA methylation in defining cellular identity and cell-type-specific gene expression [433]. However, these inhibitors act on DNA methylation across the genome and do not provide evidence for the causal role of methylation in specific regions or specific genes. Moreover 5-azacytidine was reported to have toxic effects unrelated to DNA methylation[280]. Antisense[358], siRNA[285] and gene knockout[40] depletions of DNA methyltransferases (DNMTs) provided further evidence for the role of DNA methylation in cellular differentiation and development, however DNMT depletion similarly reduces

methylation in a general manner, leaving unanswered questions as to the relative role of DNA methylation at specific regions. Furthermore, all DNMTs form complexes with chromatin silencing proteins and might control gene expression by DNA methylation independent mechanisms[423,434-436].

A study examining the state of methylation of TSS regions that are physically engaged in transcription using ChIP-sequencing with antibody against RNAPolII-PS5, the form of RNApolII that is engaged at transcription turn on, showed that promoters that are actively engaged in transcription onset are devoid of methylation[437]. Although these data show that transcription initiation is inconsistent with DNA methylation, the question of causality remains: is DNA demethylation a cause or effect of transcription onset? Similarly, enhancers are demethylated at transcription factor binding sites; is demethylation a cause or effect of transcription factor binding[438-440]?
To address this longstanding question, CRISPR/Cas9 fusion constructs with TET catalytic domains were generated to target demethylation to specific regions and to determine whether demethylation of particular regions alters transcription activity [329,330,332].

Here, we show that while dCas9-TET induces only modest demethylation of the TSS, it induces robust activation of the *Il33-002* gene (**Figure 1**), but the results leave us with unanswered questions on whether DNA demethylation of the basal promoter was causal to this activation. First, TET enzymes are not enzymatically demethylases but monooxygenases which oxidize 5-methylcytosine to 5-hydroxymethylcytosine, 5-formylcytosine, and 5-carboxylcytosine, which have demonstrated stability[241,242], demonstrated differential protein interactors[360-365], and demonstrated structural effects on DNA[366], suggesting that each derivative may be a unique epigenetic mark that confounds conclusions concerning the causality of DNA demethylation events[241,242,360-362,364].

We show here that dCas9-TET causes hydroxymethylation of the *Il33-002* promoter that is maintained in culture (**Supplementary Figure 1I**). Moreover, TET proteins are also able to oxidize thymine to 5-hydroxymethyluracil, thereby introducing another confounding epigenetic mark that produces a unique spectrum of modifications on chromatin structure and transcription factor activity[441]. A recent candidate for the improvement of such a strategy is the fusion of dCas9 to the *Arabidopsis* ROS1 glycosylase that directly removes 5-methylcytosine by direct base excision repair, foregoing the intermediate oxidized derivatives with epigenetic potential[442]; yet the issue of the overexpression of an enzyme with a capacity for unwanted and non-targeted effects is not solved by this approach.

Moreover, our data suggest that TET activation of *Il33-002* is independent of DNA demethylation since a dCas9-deadTET mutant with inhibited catalytic monoxygenase activity does not trigger demethylation but also activates *Il33-002* to a similar extent as the catalytically active dCas9-TET (**Figure 1H**). We also find that the TET2 catalytic domain is capable of inducing unmethylated DNA (**Figure 1J**), clearly indicating a demethylation-independent transactivation capacity. It is indeed known that even the restricted catalytic TET domains used in dCas9-TET fusions retain a protein interaction domain that binds O-linked N-acetylglucosamine transferase (OGT)[367,368] and TET proteins and OGT have been shown to co-localize across the genome[443]. The recruited OGT regulates gene expression by glycosylating and modulating the activity of transcription factors such as HCFC1, SP1, OCT4, MYC, p53, and RNA polymerase II as well as histones to directly increase local H2B mono-ubiquitination and trimethylation of histone 3 on lysine 4, both of which are associated with increased gene expression[206,367,443]. This mechanism as well as other potential mechanisms of catalytic-independent transcriptional activation by TET may explain our observation that dCas9-deadTET led to substantial gene induction despite an apparent lack of catalytic activity. The fact that catalytically dead TET protein activates transcription is consistent with

previous reports[63] and confounds the interpretation of the causal role of TET induced demethylation in gene activation.

Third, the fact that an enzyme with such a potential for transcriptional modulation is being overexpressed as a dCas9-TET1 fusion introduces capacity for unwanted transcriptional changes, and more recent attempts to use the SunTag system to amplify TET binding at a desired locus[330] only aggravate this issue by overexpression of large numbers of antibody-fused TET1. These undesirable effects would only be negligible in a scenario where a cell expresses a single copy of dCas9-TET that is bound at the intended locus, with highly effective oxidation and base excision repair, an impossible situation given that these lowly-active fusions must be highly expressed to facilitate robust demethylation, and thus inevitably leaving many unbound copies of dCas9-TET free to affect the genome in a TET-dependent – rather than dCas9-dependent – binding manner. Indeed, our data suggest that dCas9-TET demethylates the *Il33*-002 promoter with a scrambled, non-targeting guide (gRNAscr) (**Figure 1E-G**). We also show global genomic hypomethylation in response to dCas9-TET expression and see that a significant fraction of this demethylation localizes to enhancers, a well-established target of TET proteins[394-399]. Though these data represents only a small fraction of dmCpGs caused by dCas9-TET, enhancers are also likely a fraction of TET targets, an issue likely aggravated by both the fact that these experiments involved TET of human origin expressed in mouse cells and that the database of mouse enhancers does not necessarily reflect those to which TET is recruited in NIH-3T3 cells. This is indicative of a potential ubiquitous and dCas9-independent activity of the fused, over-expressed TET domain in a behavior similar to the demonstrated global methylation by DNMT3A in dCas9-methyltransferase fusions[316].

A second category of confounding off-target effects that are introduced by a targeting strategy that employs a flexibly tethered enzyme which can modify genetic regions in close physical proximity – despite large genetic distances – particularly those in

ubiquitous self-interacting topologically associating domains (TADs), such as the one to which the two *Il33* promoters belong[444]. This is aggravated by the fact that the TET family of proteins is known to participate in enhancer regions and facilitate long-range chromatin interactions[445,446]. However, as mentioned above, this may also not be a long-range interaction, but rather a direct interaction of dCas9-TET through the TET domain. Thus, whenever a flexibly tethered enzyme is employed for epigenetic editing, it will be difficult to dissociate effects of targeted and nontargeted DNA demethylation on transcription activity.

Finally, the demethylation that is observed with dCas9-TET fusions might be secondary to transcription activation. When we combined our three targeting gRNAs with the well-characterized dCas9-VP64 fusion, VP64 is a potent transcriptional activator originating from the herpes simplex virus[305], we observed broad demethylation of the *Il33-002* promoter (**Supplementary Figure 1F-H**). This phenomenon suggests that DNA demethylation can in particular instances be secondary to transcription factor recruitment and transcriptional activation (**Figure S1I-J**) as has been previously reported [438,439].

Lastly, there are examples in which dCas9 or another targeting protein either bears a catalytically inactive form of TET or the domain is altogether missing, and mild demethylation is still observed[311,329,330,447]. We propose that, in some cases, this demethylation stems from the lingering transactivation capacity of the mutated TET domain (discussed above) followed by demethylation as a consequence of activation, such as the demethylation caused by VP64 activation (**Figure S1F-H**). Alternatively, as we demonstrate here (**Figure 2**) binding of dCas9 blocks DNA methyltransferase catalyzed methylation. This therefore obscures the true contribution of TET proteins to demethylation. It is in fact possible that most of the demethylation triggered by dCas9-TET fusions seen in dividing cells stems from the simple steric interference with DNA methyltransferase activity as we demonstrate in this study.

Taken together, these data reveal that while dCas9-TET may be a valid tool for producing epigenetic perturbations that may further understanding of TET dynamics, it introduces a number of confounds inherent to the properties of the TET protein that prohibit conclusions as to the causal relationship of changes in DNA methylation at particular sites and gene expression.

We instead propose and demonstrate here a previously unrecognized capacity of dCas9 to prevent DNA methylation with high efficacy at fairly small, precise regions and, more importantly, free from any fused eukaryotic enzyme that may act independently of the dCas9:gRNA binding activity. We first show that this approach can be implemented to map the individual methylated CpGs within a regulatory region which silence transcription using an *in vitro* methylation promoter-reporter transient-transfection assay. This method has advantages over earlier methods that protected individual CpGs from methylation by mutagenesis to non-CpG sequences[440], since mutagenesis can disrupt protein:DNA interactions by the sequence change rather than by the methylation difference[448]. Our method alters the methylation per se without disrupting the genetic sequence. Our results demonstrate that three CpG sites within 22 bp of the TSS are sufficient to silence the *Il33-002* promoter while other CpG sites do not contribute to methylation dependent silencing of promoter activity.

We further show that this approach can be applied to trigger site-specific demethylation in dividing cells and that it can be optimized for near-complete removal of DNA methylation from sites that had previously been fully methylated, without perturbing the methylation states of adjacent CpGs in the same promoter to any substantial degree. Thus, this method could interrogate the causal role of DNA methylation in silencing gene expression. Since inhibition of DNA methylation is dependent on tight binding of dCas9 which is also dependent on gRNA target and quality, the risk for nontargeted demethylation is low. Accordingly, we find that there appear to be no off-target DNA

demethylation events as a consequence of gRNA:dCas9 off-target binding in WGBS/ChIP-seq data and in targeted sequencing of 5 candidate off-target regions. However, further work is needed to identify the biological origin of the dmCpGs that are not a consequence of dCas9:gRNA off-target binding that were detected in WGBS analysis. Potential off-target effects of a larger number of gRNAs across multiple cell lines and species need to be evaluated as well.

We used our method of demethylation to define the role of TSS and proximal promoter methylation of the *Il33-002* gene in its cogent genomic context. We found that demethylation of the *Il33-002* TSS produces a small but significant increase in its expression. Our results confirm what was observed in the transient transfection assay: CpG sites 9 to 11 at the TSS suppress promoter activity. However, dCas9-TET induced 25-fold higher *Il33* expression compared to dCas9 alone when targeted to the same promoter, even though it caused significantly lower demethylation than dCas9 (**Figure 5H**). There are several possible explanations for this discrepancy between the fold induction achieved by demethylation and by TET recruitment. First, the fusion of TET to dCas9 is flexible and may allow access to DNA in a wider region, perhaps inducing demethylation in other regulatory regions or methylated transcription factors that are required for more robust expression (**Figure 1 E-G**). However, treating cells that have been demethylated at the *Il33-002* TSS CpG sites 9-11 with 5-2' deoxy-azacytidine doesn't further induce the gene, while cells that were methylated at 9-11 sites are induced to a level like the levels achieved by dCas9. This suggests that the main regulation by DNA methylation occurs at CpGs 9-11 but that the gene is further induced by DNA methylation independent mechanisms that are partially triggered by TET. This illustrates that the results of TET targeting could not be automatically understood as being driven by demethylation and highlights the need for enzyme independent targeted DNA demethylation for understanding the role of DNA methylation.

We then determined whether demethylation of the TSS poises the *Il33-002* promoter for induction by known inducers of this gene. poly(I:C) induces *Il33-002* 300-fold without detectable DNA demethylation and does not induce the demethylated *Il33-002* to a higher level than the methylated *Il33-002* promoter. Thus, induction of *Il33-002* expression is independent of DNA demethylation in the basal promoter. It is possible however that poly(I:C) triggers demethylation in a remote enhancer that wasn't examined in our study. In contrast, induction by LPS is higher when the basal promoter is demethylated, however LPS induces the promoter whether it is methylated or not suggesting an additive but nonessential effect of demethylation of TSS for LPS induction. What is the role of *Il33-002* promoter methylation? The data is consistent with the idea that this gene is mainly regulated by extracellular signals irrespective of DNA methylation. DNA methylation only silences the residual basal activity of the promoter, perhaps to prevent leaky expression and transcriptional noise in the absence of the appropriate signal. This is consistent with the observation that the ectopically transfected promoter is silenced by DNA methylation (**Figure 2**). Therefore, either targeted demethylation or 5-aza-2'-deoxycytidine achieve only a small elevation in expression.

A different paradigm is represented by the *SERPINB5* promoter. Demethylation of the basal promoter on its own has no effect on expression, which remains low (**Figure 7E-F**) even when 6 CpGs in the proximal promoter region become completely demethylated (**Figure 7 B-E**). However, global demethylation by 5-aza -2'-deoxycytidine induces the activity of this demethylated promoter further than the naturally methylated promoter in control MDA-MB-231 cells, suggesting that expression of this gene is regulated by methylation in the promoter region as well as other regions in *cis* or *trans*. Demethylation of the proximal promoter on its own is insufficient to induce transcription. A possible explanation is that activity of this gene requires activation of a transcription factor that is silenced in these cells and induced by demethylation as we have recently shown[449]. The tumor necrosis factor (*Tnf*) gene exhibits a proximal TSS promoter region

that is highly methylated in NIH-3T3 cells. The gene is highly induced and its proximal promoter region is demethylated by 5-aza-2'-deoxycytidine (**Figure 7G**). In contrast to the large induction of expression of by 5-aza-2'-deoxycytidine, demethylation of 10 CGs proximal to the TSS (**Supplementary Figure 9**) using the targeted dCas9 method did not turn on the gene (**Figure 7H**, grey bars). Here, as was the case with the *SERPINB5* promoter, 5-aza-2'-deoxycytidine treatment of cells bearing dCas9-demethylated *Tnf* TSS region resulted in higher induction of expression than treated control cells bearing a methylated *Tnf* TSS (**Figure 7H**). These experiments illustrate the importance of studying demethylation of specific sites *per se* to truly understand their contribution to gene expression control.

Finally, in a manner dissimilar to the other genes examined in this study, targeted demethylation of the large, highly-methylated *FMR1* repeat region in Fragile X syndrome patient fibroblasts did induce basal transcription of the *FMR1* gene up to a 110-fold in one cell pool suggesting that, in this case, methylation of the repeat element plays a large role in silencing of the gene.

As the larger magnitude of demethylation observed in the dCas9 approach does not produce transcriptional changes as substantial as those observed by dCas9 tethered to TET1, it is clear that promiscuous mammalian enzymatic domains do not exclusively demethylate, have other methylation independent activities, and cannot be suitably applied to investigate the causal relationship between DNA methylation at specific sites and gene expression.

The ability of newly demethylated sites to stay demethylated can vary; we detected no increase in *Il33-002* TSS methylation 75 days after removing dCas9 by Cre-mediated recombination in NIH-3T3 cells but saw a small non-significant increase in SERPINB5 demethylated CpGs 45 days after Cre-mediated dCas9 removal. It is useful for research timescales that sites stay demethylated, but the small variation between the two genes in the two cell lines suggests that the retention of unmethylated CpGs may vary as a

factor of cell line (e.g. how much de novo methyltransferase activity a cell line has) or by specific CpG sites (e.g. in a growing cell population, how detrimental to cell growth is demethylation of a specific CpG, and will it be selected against?) and thus it will be important in future studies that use this technique to assay how long demethylation persists in the CpG and cell line contexts under examination to ensure that demethylation persists for the duration of the experiments. It is important to note that in stem cells where de novo DNMTs are expressed to a higher level[325], methylation might be regained after removal of dCas9.

In summary, we developed a tool that allows site-specific demethylation of a narrow region of DNA by physical blocking of DNMTs without using confounding epigenetic enzymatic activities. This tool enables the examination of causal relationships between demethylation of specific sites and gene expression in genes at their native positions in the chromatin. Comparing the results obtained using this tool and results obtained using general DNA methylation inhibitors reveals that the role of DNA demethylation at specific sites might have been previously overestimated by confounded techniques, and thus is part of a growing body of evidence in support of this notion[203,450]. Our study demonstrates the need for the careful causational investigation of the role of DNA demethylation of different regions *per se* by an unconfounded tool. We hope that this tool can be used to attribute causality to DNA methylation changes not only in fundamental physiological gene transcription, but also under different specific physiological and pathological conditions mediated by changes in extracellular signals and changes in the milieu of cellular transcription factors in order to begin to reveal the true extent, the nature, and the diverse contribution of DNA methylation at different regions to gene regulation.

12.5 Methods

gRNA design and synthesis

To maximize likelihood of on-target efficiency and minimize off-target binding, gRNAs were designed using three online tools with distinct scoring algorithms: Off-Spotter, CCTOP, and CRISPR Design[378,387,388]. Final gRNAs were chosen based on highest cumulative rank and location in the promoter. The scrambled gRNA sequence was obtained from pCas-Scramble (Origene). For *in vitro* assays, gRNAs were *in vitro* transcribed with the GeneArt™ Precision gRNA Synthesis Kit (Thermo Fisher Scientific) according to manufacturer protocol and using primers listed in **Supplementary Table 2.** Due to a lack of available kit compatible with *S. aureus* gRNAs, SA-gRNA1-4 and SP-gRNA1-4 were generated by a custom T7 *in vitro* transcription protocol (dx.doi.org/10.17504/protocols.io.dwr7d5) modified to replace the *S. pyogenes* scaffold sequence with that of *S. aureus*. (primers in **Supplementary Table 3**). Lentiviral gRNAs were first produced according to the protocol by Prashant Mali[451]. Briefly, 455bp double stranded DNAs containing the human U6 promoter, gRNA sequence, gRNA scaffold, and termination signal were ordered as gBlock Gene Fragments (IDT). These were re-suspended, amplified with Taq Polymerase (Thermo Fisher Scientific) according to manufacturer protocol and using primers listed in **Supplementary Table 2**, extracted from an agarose gel with the QIAEX II Gel Extraction Kit (QIAGEN), and inserted into pCR®2.1-TOPO (Thermo Fisher Scientific) by incubating for 30 minutes at room temperature. The gRNA scaffold was now flanked by EcoRI sites from the vector. A lentiviral backbone was obtained from Addgene (pLenti-puro, Addgene #39481) and the CMV promoter was removed to prevent aberrant transcription by digesting the plasmid with ClaI-HF and BamHI-HF (NEB), gel extracting, removing DNA overhangs with the Quick Blunting™ Kit (NEB), and circularization with T4 Ligase (Thermo Fisher Scientific) for 1 hour at 22°C. The resulting promoterless pLenti-puro plasmid was then digested with EcoRI and the 5' phosphates were removed with Calf Intestinal Alkaline Phosphatase (Thermo Fisher Scientific) to facilitate efficient ligation of the EcoRI-flanked gRNA scaffold. Resulting clones were Sanger sequenced with pBABE 3' sequencing primer to ensure proper gRNA sequence and orientation (Génome Québec). The gRNAs targeting *SERPINB5* and *Tnf* were created by site-directed

mutagenesis of pLenti-Il33gRNA6-puro using primers listed in **Supplementary Table 2** and the Q5® Site-Directed Mutagenesis Kit (NEB) according to manufacturer protocol. The *HNF4A*-targeting gRNA is from the genome-scale CRISPR knock-out (GeCKO) v2 library[452] (purchased as lentiviral plasmid from Genscript) and the FMR1-targeting gRNA from the Jaenisch lab was obtained from Addgene (pgRNA-CGG, Addgene #108248).

Site-specific *in vitro* DNA methylation

First, a dCas9:gRNA ribonucleoprotein complex was formed with the following mixture: 14 µL nuclease-free water, 3 µL Cas9 Reaction Buffer (Applied Biological Materials Inc.), 7.5 µL 300 nM CpG-targeting *in vitro* transcribed gRNA or non-CpG-targeting control gRNA, and dCas9 recombinant protein (Applied Biological Materials Inc.). After 10 minutes at room temperature, 3 µL of 30 nM Il33-pCpGl was added to the reaction, which was then transferred to 37°C to allow dCas9:gRNA complex binding to DNA. After 1 hour, the following mixture was added to the reaction: 145 µL nuclease-free water, 17 µL NEBuffer™ 2, 5 µL 32mM S-Adenosyl methionine (NEB) (final concentration 0.8 mM) , and 3 µL (12 units) M.SssI methyltransferase (NEB). This solution was pre-warmed to 37°C before addition to prevent interference with dCas9:gRNA binding to the DNA. After 4 hours of incubation at 37°C, 1 µL of 20 mg/mL Proteinase K (Roche) was added and the temperature was raised to 64°C for an additional 4 hours.

DNA Isolation, bisulfite conversion, bisulfite-cloning, and pyrosequencing

Plasmid DNA was recovered by phenol-chloroform extraction and precipitation in ethanol overnight. DNA was washed one time with 70% ethanol, dried, and re-suspended in 30 µL nuclease-free water. Genomic DNA was extracted from cells by resuspension in 400 µL DNA lysis buffer (100mM Tris, pH 7.5, 150 mM NaCl, 0.5%

SDS, 10 mM EDTA) and treatment with 2 µL 20mg/mL RNAse A (NEB) for 30 minutes at 37°C and 5 µL 20 mg/mL Proteinase K (Sigma) for 4 hours at 55°C. This was followed by phenol-chloroform extraction by addition of 200 µL phenol solution and 200 µL of chloroform, vortexing for 10 seconds, and centrifugation at 16,000 xg for 5 minutes at 4°C. The aqueous phase was then transferred to a new 1.5 mL tube, mixed with 400 µL chloroform, and centrifuged again at 16,000 xg for 5 minutes at 4°C. The aqueous phase was again transferred to a new tube and DNA was precipitated by the addition of 1 mL 95% ethanol and 1 µL glycogen overnight at -80 °C, centrifugation at 16,000 xg for 30 minutes at 4°C. DNA was washed a single time with 1 mL 70% ethanol, centrifuged at 16,000 xg for 15 minutes at 4°C, air dried for 5 minutes, and resuspended in 50 µL nuclease-free water. Following DNA extraction, bisulfite conversion was conducted according to manufacturer protocol with the EZ DNA Methylation-Gold Kit (Zymo Research) using 5 µL of *in vitro* methylated plasmid DNA or 1.5 µg genomic DNA measured with the Qubit dsDNA BR Assay (Thermo Fisher Scientific). 1 µL of bisulfite-converted DNA was amplified with HotStar Taq DNA polymerase (QIAGEN) in a 25 µL reaction using the primers designed with MethPrimer [453] and listed in **Supplementary Table 2**. Pyrosequencing samples were processed in the PyroMark Q24 instrument according to protocols designed by the PyroMark Q24 software (QIAGEN). Sequencing primers were designed with Primer3 [454]. Alternatively, amplicons were cloned into pCR®4-TOPO (Thermo Fisher Scientific) for 30 minutes at room temperature and transformed into TOP10 competent cells (Thermo Fisher Scientific) prior to plasmid isolation with the High-Speed Plasmid Mini Kit (Geneaid) and Sanger sequencing (Eurofins Genomics) using the M13R sequencing primer. All oligonucleotides used in this study were purchased from Integrated DNA Technologies.

Luciferase Assay

$8.0x10^4$ NIH-3T3 cells (*Il33* experiments) or $1.2x10^5$ HEK293 cells (TET co-transfection) were plated in a 6-well plate (Corning) 24 hours prior to transfection. 1 µg (*Il33*) or 100

ng (SV40) plasmid DNA from the *in vitro* methylation reactions were transfected with 3 µL (*Il33)* or 1 µL (SV40) X-tremeGENE 9 transfection reagent (Roche) diluted in 50 µL of Opti-MEM medium (Gibco). Luciferase assays were performed 36 hours after transfection using the Luciferase Reporter Gene Assay, high sensitivity (Roche). Briefly, cells were washed with 1 mL of phosphate-buffered saline (Wisent), detached with scrapers (Thermo Fisher Scientific) after the addition of 150 µL lysis buffer, and transferred to 1.5 mL tubes. After a 15-minute incubation at room temperature, the mixtures were centrifuged for 5 seconds at maximum speed and the supernatant transferred to new 1.5 mL tubes. Two 50 µL volumes per condition were supplemented with 50 µL luciferase assay reagent in disposable glass tubes (Thermo Fisher Scientific) and light emission was measured immediately in the Monolight 3010 luminometer (Analytical Luminescence Laboratory). Sample protein concentration was determined by Bradford Protein Assay (Bio-Rad) and A595 readings were measured in a DU 730 UV-Vis Spectrophotometer (Beckman Coulter). Protein concentration in cell lysate was determined by comparing to a bovine serum albumin standard curve and luciferase activity was normalized to concentration. We validated that our transfection method results in equal copy numbers transfected for both methylated and unmethylated DNA by measuring copy number of transfected pCpGl 36 hours after transfection (**Supplementary Figure 1N**).

Plasmids

The original dCas9 plasmid lacking loxP sites was obtained as a dCas9-VP64 fusion (lenti dCAS-VP64_Blast, Addgene #61425). The VP64 domain was removed by digestion with BamHI-HF and BsrGI-HF, blunting with the Quick Blunting™ Kit (NEB), and circularization with T4 Ligase (Thermo Fisher Scientific) for 1 hour at 22°C. Following transformation, plasmids were isolated from ampicillin-resistant clones (High-Speed Plasmid Mini Kit, Geneaid) and Sanger sequenced to identify plasmids that maintained the blasticidin resistance gene in-frame with dCas9. Floxed dCas9 was

purchased as a ready plasmid (pLV hUbC-dCas9-T2A-GFP, Addgene #53191) and primers were designed to amplify a fragment of approximately 500 base pairs when dCas9 is removed with Cre recombinase (**Supplementary Table 2**). The Cre-containing plasmid was obtained from Addgene (pLM-CMV-R-Cre, Addgene #27546). A fragment encoding the CMV promoter and mCherry-T2A-Cre-WPRE was excised by NdeI and SacII (Thermo Fisher Scientific) and transferred to the pLenti6/V5-DEST™ Gateway™ Vector (Thermo Fisher Scientific) bearing a blasticidin resistance cassette (Thermo Fisher Scientific) to facilitate antibiotic selection. Lentiviral Fuw-dCas9-Tet1CD-P2A-BFP and Fuw-dCas9-dead Tet1CD-P2A-BFP were obtained from Addgene (Addgene #108245,#108246). Catalytically active Cas9 lentiviral vector was obtained from Genscript as pLentiCas9-Blast. TET1 plasmids were obtained from Addgene: #49792 (FH-TET1-pEF) and #124081 (pEF1a_FL MUT TET1) and control pEF1A was purchased from Thermo Fisher Scientific. pcDNA3-TET2 (Fig S1J) was generated by amplification of TET2 from human cDNA, TOPO-TA cloning and sequence validation by Sanger sequencing, followed by digestion and ligation into pcDNA3.1 (Thermo Fisher Scientific) using the restriction enzymes XhoI and ApaI. SV40-pCpGl (Fig 1J) was generated by amplification of the SV40 promoter and enhancer region from lenti dCAS-VP64_Blast using primers that added a 5' BamHI site and a 3' HindIII site, which were then used for transfer into pCpGl[299] following sequence verification.

Cell culture

HEK293T and NIH-3T3 cells (ATCC) were thawed and maintained in DMEM medium (Gibco) supplemented with 10% Premium Fetal Bovine Serum (Wisent) and 1X Penicillin-Streptomycin-Glutamine (Gibco). Cells were grown in a humidified incubator of 5% carbon dioxide at 37ºC and cultured in 100mmx20mm tissue culture dishes (Corning) and harvested or passaged by trypsinization (Gibco) upon reaching 80-90% confluency. Clones were isolated by limiting dilution and trypsinization with the aid of cloning rings. Fragile X syndrome fibroblasts (GM05848, Coriell Institute) and age-matched control fibroblasts (GM00357, Coriell Institute) were maintained as above. Flow cytometry to isolate dCas9-TET/dCas9-deadTET (BFP) and dCas9 (GFP) when

antibiotic selection was not an option was performed by Julien Leconte of the Flow Cytometry Core Facility at McGill University Life Sciences complex. All replicates presented in this study are biological replicates. A technical replicate is performed for each assay and averaged per each biological replicate.

Lentiviral production

HEK293T cells were plated at a density of 3.8 x 10^6 per 100mm dish 24 hours prior to transfection. Cells were transfected using X-tremeGENE 9 transfection reagent (Roche). Briefly, individual lentiviral transfer plasmids were mixed with a packaging plasmid (pMDLg/pRRE, Addgene #12251), envelope protein plasmid (pMD2.G, Addgene #12259), REV-expressing plasmid (pRSV-Rev, Addgene #12253), and the transfection reagent in Opti-MEM medium (Gibco). The mixture was incubated for 30 minutes at room temperature and added in a drop-wise manner to HEK293T cells in 8 mL of fresh DMEM medium in a 100mm dish. Lentiviral particles were harvested by filtering the supernatant through a 0.45 µm disk filter 72 hours after transfection and either used immediately or stored at -80ºC. 5 µg/mL Blasticidin S HCl and 1-20 µg/mL Puromycin Dihydrochloride (Gibco) were used to select for stable transformants.

Transient transfection for time-course experiments

8.0x10^4 NIH-3T3 cells stably expressing dCas9-VP64 (from lentiviral transfer and blasticidin selection, above) were plated in a 6-well plate (Corning) 24 hours prior to transfection. 1 µg per well pLenti-IL33_gRNA2 vector was transfected using X-tremeGENE 9 transfection reagent (Roche) and cells were harvested at 0, 24, 48, 72, and 96 hours. RNA and DNA were extracted from separate wells and RNA expression and DNA methylation were measured as described in the relevant methodology sections.

RT-qPCR

RNA was isolated from approximately 80% confluent 100mm dishes with 1 mL of Trizol reagent (Thermo Fisher Scientific) following harvest by trypsinization and washing with phosphate-buffered saline (Wisent). RNA extraction was performed according to Trizol manufacturer protocol. Briefly, 200 mL of chloroform was added to 1 mL of Trizol-RNA mixture. The samples were thoroughly vortexed, incubated at room temperature for 2 minutes, and centrifuged for 15 minutes at 12,000 xg at 4°C. The aqueous phase was transferred to a new 1.5 mL tube prior to the addition of 0.5 mL isopropanol and incubation at room temperature for 10 minutes. The samples were centrifuged for 10 minutes at 12,000 xg at 4°C, and washed twice with 75% ethanol, discarding the supernatant each time. The pellets were air dried for 10 minutes and re-suspended in 50 µL DEPC-treated water (Ambion). Concentrations were measured with the Qubit RNA BR Assay (Thermo Fisher Scientific) and 1 µg RNA was used for each reverse transcriptase reaction using M-MuLV Reverse Transcriptase (NEB) according to manufacturer protocol. cDNA was diluted 1:2 (20 µL reverse transcription reaction to 40 µL water) and 2 µL of diluted cDNA was amplified in the LightCycler ® 480 Instrument II (Roche) in a 20 µL reaction containing 10 µL LightCycler ® 480 SYBR Green I Master Mix (Roche) and 0.8 µL each of 10 µM forward and reverse primer listed in **Supplementary Table 2**. Quantification was performed by Roche Lightcycler Software.

Drug treatment

5-aza-2'-deoxycytidine (Sigma A3656) was dissolved to 10 mM in sterile water and frozen in one-time-use aliquots at -80°C. Trichostatin A (TSA, Sigma T8552) was dissolved to 1 mM in dimethyl sulfoxide (DMSO, Sigma D8418) and frozen in one-time-use aliquots at -80°C. Lipopolysaccharides from Escherichia coli O55:B5 (Sigma L6529) were diluted to 1mM in phosphate-buffered saline. 5-aza-2'-deoxycytidine and TSA

treatment regimen involved 3 treatments every other day with media replacement (5 days total) at specified concentrations and sample collection on the sixth day.

Off-target prediction for pyrosequencing

Potential off-target sites of *Il33* gRNA3 in the mouse genome were predicted using Cas-OFFinder[389], a program that allows bulges in the RNA and DNA (which Cas9 is known to tolerate) to increase the number of possible off-target sites. Because we were interested in changes in methylation, results were filtered for the presence of a CG at a maximum of 10bp from either end of the gRNA sequence. Of 15 results, 2 differed by 3 mismatches, 9 by 4 mismatches, and 4 by 2 mismatches and a bulge. We developed functional pyrosequencing assays for 4 of these sites.

Hydroxymethylation quantification

DNA isolated from cells by phenol:chloroform isolation and ethanol precipitation was cleaned on Micro Bio-Spin P-6 SSC columns (Bio-Rad) according to manufacturer protocol. 15 mM $KRuO_4$ (Sigma) was prepared by dissolving 0.153g in 50 mL of 0.05M NaOH and thawed freshly for each oxidation reaction. 1 µg cleaned DNA was incubated in a 19 µL volume reaction in a PCR tube with 0.95 µL 1M NaOH at 37 °C in a shaking incubator for 0.5 hr. The sample was cooled immediately in an ice-water bath for 5 min prior to the addition of 1 µL ice-cold 15 mM $KRuO_4$ and incubation in an ice-water bath for 1 hr with vortexing every 20 min. A second oxidation was performed by the addition of 4 µL 0.05 M NaOH, incubation at 37 °C in a shaking incubator for 0.5 hr, following by cooling, addition of 1 µL ice-cold 15 mM KRuO4, and incubation in ice-water bath with occasional vortexing as before. Oxidized DNA was cleaned again on Micro Bio-Spin P-6 SSC columns and the DNA was subjected to bisulfite conversion and pyrosequencing. Control reactions were done in parallel in which 15 mM KRuO4 was replaced by 0.05 M

NaOH and percent hydroxymethylation was quantified as the decrease in methylated fraction in oxidized DNA as compared to control DNA.

Chromatin immunoprecipitation

150mm tissue culture dishes containing 90% confluent NIH-3T3 cells from each experimental condition were cross-linked by the direct addition of formaldehyde to a 1% final concentration. The dishes were incubated for 10 minutes at room temperature with constant agitation. The reaction was quenched by the addition of glycine to a final concentration of 0.125 M and incubated for an additional 5 minutes at room temperature with constant agitation. Cross-linking solution was aspirated and cross-linked cells were washed three times with 10 mL ice-cold phosphate-buffered saline (PBS). 10 mL of ice-cold PBS was added and cells were scraped into suspension by a rubber cell scraper. Cross-linked cells were pelleted at 800xg at 4°C in 15mL falcon tubes, the supernatant removed, and the cells were lysed in 300 µL ice-cold lysis buffer (1% SDS, 10 mM EDTA, 50 mM Tris), pipeted thoroughly, incubated for 15 minutes on ice, and immediately sonicated on the Bioruptor (Diagenode) in 1.5 mL Eppendorf tubes at high power for three 10-minute cycles of 30 seconds on and 30 seconds off, replacing warmed water with ice-cold water and minimal ice between each cycle. Sonicated samples were centrifuged at 4°C for 16,000xg for 5 minutes, and supernatant was transferred to a clean 1.5 mL Eppendorf tube, with 30 µL set aside for shearing efficiency analysis. The remaining supernatant was diluted with a 9X volume of dilution buffer (16.7mM TrisHCl pH 8.0, 1.2mM EDTA, 167mM NaCl, 0.01% SDS, 1.1% Triton X-100) and precleared with washed Dynabeads Protein G (Thermo Fisher) for 2 hrs at 4°C on a nutator. Using a magnetic rack, 1% of pre-cleared chromatin was set aside for input and 5 µg Monoclonal ANTI-FLAG® M2 antibody (Sigma, F1804) (to capture 5' 3xFLAG-tagged dCas9) or IgG (abcam) was added to the remaining volume and then incubated at 4°C on a nutator overnight. 60 µL of washed (3X with Tris-EDTA – 10mM Tris pH=8, 1mM EDTA – and 3X with RIPA – 20 mM Tris, 2 mM EDTA, 150 mM NaCl,

1% Triton X, 0.1% SDS, 0.5% deoxycholate) Dynabeads were added to each sample and incubated at 4°C on a nutator for 4 hrs. Beads were then washed with 1mL each as follows: 2X with low salt wash buffer (0.1% SDS, 1% Triton-X, 2mM EDTA, 20mM Tris, 150mM NaCl), 2X with high salt wash buffer (same as low except 500 mM NaCl), 2X with LiCl wash buffer (0.25M LiCl, 1% NP-40, 1% deoxycholate, 1mM EDTA, 10mM Tris, pH 8.0), and 2X with Tris-EDTA. All buffers contained 1X cOmplete™ Protease Inhibitor Cocktail (Sigma). DNA was eluted by the addition of 100 µL elution buffer (1% SDS, 0.1M $NaHCO_3$), vortexing vigorously, and 15-minute incubation at room temperature with constant agitation before transferring to a clean 1.5 mL tube. This was repeated twice for a final volume of 200 µL and the input fraction was adjusted to the same volume with elution buffer. Reverse cross-linking (0.2M final concentration of NaCl, 65 °C overnight) was performed for all samples, followed by standard treatment with RNAse A, proteinase K, and phenol:chloroform cleanup followed by ethanol precipitation. Clean DNA was then quantified by qPCR and enrichment in the immunoprecipitated samples was calculated as fraction of input. Nonspecific (IgG) antibody and qPCR primers of unbound regions were used as controls for effective immunoprecipitation.

Chromatin immunoprecipitation sequencing and analysis

For ChIP-seq experiments, cells were prepared as for ChIP and IP was performed with the same anti-FLAG antibody (above) on NIH-3T3 expressing FLAG-tagged dCas9-GFP (selected by FACS) and gRNA (selected under high puromycin); these are the same cells depicted in **Figure 5** (transduced with empty vector instead of Cre recombinase). All cross-linking and immunoprecipitation steps were performed with the ChIP-IT High Sensitivity® Kit (Active Motif) according to manufacturer's instructions using 30 µg input chromatin as quantified by NanoDrop. Sonication was performed as above. Successful ChIP with anti-FLAG antibody was validated by qPCR (as described above) with primers for *Il33* (positive control) and *Actb* (negative control)

(**Supplementary Figure 11**). Eluted DNAs were sent to Centre d'expertise et de services Génome Québec at McGill University for library preparation and sequencing. Fragmented DNA from 12 samples (three replicates each of gRNAscr anti-FLAG, gRNA3 anti-FLAG, gRNAscr input, and gRNA3 input) was quantified using 2100 Bioanalyzer (Agilent Technologies). Libraries were generated robotically with fragmented DNA (range 100-300 bp) using the NEBNext Ultra II DNA Library Prep Kit for Illumina (New England BioLabs), as per the manufacturer's recommendations. Adapters and PCR primers were purchased from Integrated DNA Technologies (IDT). Size selection was carried out using SparQ beads (Qiagen) prior to PCR amplification (12 cycles). Libraries were quantified using the Kapa Illumina GA with Revised Primers-SYBR Fast Universal kit (Kapa Biosystems). Average size fragment was determined using a LabChip GX (PerkinElmer) instrument. The libraries were normalized and pooled and then denatured in 0.05N NaOH and neutralized using HT1 buffer. The pool was loaded at 200pM on a Illumina NovaSeq S4 lane using Xp protocol as per the manufacturer's recommendations. The run was performed for 2x100 cycles (paired-end mode). A phiX library was used as a control and mixed with libraries at 1% level. Base calling was performed with RTA v3.4.4 . Program bcl2fastq2 v2.20 was then used to demultiplex samples and generate fastq reads. Paired-end FastQ files were trimmed for adapters and quality scores using TrimGalore v0.6.4_dev[455] under default settings. Alignments to the mm10 genome were performed using bowtie2 v2.3.4.1[456] under default settings and peak calling for each sample was performed with the macs2 v2.2.7.1[457] callpeak function (--g mm --nomodel --extsize 204 --SPMR) after first running the predictd script and establishing --extsize 204 according to the macs2 manual. Alignments were passed to the DiffBind R package to identify significantly differentially enriched regions under default parameters.

Western blot

Control NIH-3T3 cells and cells expressing dCas9-TET or dCas9-deadTET were grown to 80% confluency on 100mm tissue culture dishes. Cells were washed twice with 10mL PBS and collected into 15mL falcon tubes by scraping and then pelleted by centrifugation for 5 minutes at 300xg at 4°C. The supernatant was aspirated and cells were resuspended in 80 µL protein extraction buffer (150 mM NaCl, 0.1% SDS, 0.5% sodium deoxycholate, 50 mM Tris pH 7.5, and 1% NP-40) with 1X cOmplete™ Protease Inhibitor Cocktail (Sigma), incubated for 30 minutes on ice with vortexing every 5 minutes, centrifuged for 10 minutes at 16,000 xg at 4°C. The supernatant was retained and protein concentration was measured by Bradford assay. 20 µg protein in 2x Laemmli Sample Buffer (Bio-Rad) was prepared according to manufacturer protocol and loaded into a 5% acrylamide gel (for dCas9-TET/deadTET) or 10% acrylamide gel (for beta-actin loading control) with 5% upper stacking gel. Gels were run for 10 minutes at 110 V and then for 50 minutes at 170 V, followed by overnight transfer to nitrocellulose membrane at 30V. Membranes were blocked with 1% milk in TBST and protein was detected with either mouse Anti-CRISPR-Cas9 primary antibody [7A9-3A3] (Abcam, ab191468) (1/2,000 dilution) or monoclonal Anti-β-Actin primary antibody produced in mouse (Sigma, A2228) (1/5,000 dilution)and goat Anti-Mouse IgG H&L (HRP) secondary antibody (Abcam, ab205719) (1/10,000 dilution). Each antibody incubation was performed for 1 hour. After addition of Clarity Western ECL Substrate (BioRad), images were acquired with automatic exposure on the Amersham Imager 600.

Whole-genome bisulfite sequencing (WGBS)

WGBS was performed by the Centre d'expertise et de services Génome Québec at McGill University. Genomic DNA was quantified using the Quant-iT™ PicoGreen® dsDNA Assay Kit (Life Technologies). 2x151bp paired-end libraries were generated using the NEBNext® Enzymatic Methyl-seq Kit (New England BioLabs, NEB). Adapters were purchased from NEB. Libraries were quantified using the Kapa Illumina GA with Revised Primers-SYBR Fast Universal kit (Kapa Biosystems) and average size

fragment was determined using a LabChip GX (PerkinElmer) instrument. The libraries were normalized and pooled and then denatured in 0.05N NaOH and neutralized using HT1 buffer. The pool was loaded at 225pM on an Illumina NovaSeq S4 lane using Xp protocol as per the manufacturer's recommendations. The run was performed for 2x100 cycles (paired-end mode). A phiX library was used as a control and mixed with libraries at 5% level. Base calling was performed with RTA v3.4.4 . Program bcl2fastq2 v2.20 was then used to demultiplex samples and generate FastQ reads.

WGBS data analysis

Paired-end FastQ files were trimmed for adapters and quality scores using TrimGalore v0.6.4_dev [455] under default settings. Alignments to the mouse mm10 genome, deduplication, and methylation calling were performed using Bismark v0.22.3[458] under default settings. All statistical analyses were performed with the R package methylKit v1.14.2 [459]. For off-target analyses for dCas9:gRNA3:Cre, significantly differentially methylated (q<0.01, methylation difference >25%) CpGs were determined by comparison to dCas9:gRNAscr:Cre with the calculateDiffMeth function after filtering for CpGs that were covered at least 5X in all samples. Off-target site manhattan plot generated with R package qqman[460].

Quantification of *FMR1* CGG repeat methylation

DNA from Fragile X patient fibroblasts treated with lentiviral dCas9 and either lentiviral gRNAscr or gRNA-CGG was isolated by the phenol-chloroform method as described above. 2 µg DNA from each condition was digested for 4 hours at 37 °C in a thermocycler in a 20 µL reaction containing 2 µL rCutsmart buffer and 1 µL Fnu4HI restriction enzyme (NEB) or in a control reaction without enzyme. Methylation sensitivity of the enzyme was verified in parallel by digestion of unmethylated or *in vitro* (M.SssI) methylated plasmid DNA and agarose gel electrophoresis. Following restriction digest,

DNAs were re-purified using Monarch® PCR & DNA Cleanup Kit (NEB). DNA concentration was measured by NanoDrop and DNAs were diluted to 20 ng/μL for use with the AmplideX® mPCR FMR1 assay (Asuragen). Note that the AmplideX® mPCR FMR1 assay involves restriction digest with methylation sensitive enzyme HpaII that is directly outside the CGG repeat region and is not informative for the methylation status of the CpG dinucleotides that make up the CGG region. Therefore, the protocol was modified as described above to allow for digestion with the methylation sensitive enzyme Fnu4HI (recognizes GCNGC) and PCR amplification was carried out with only the control workflow (FAM: no digestion) from the manufacturer. Briefly, 8 μL of diluted sample DNA was mixed with 2 μL control DNA. 4 μL of this mixture was incubated for 2 hours at 37 °C with 3.7 μL Digestion Buffer and 0.3 μL Control Enzyme (FAM). Then 20 μL GC-Rich Amp Buffer, 0.1 μL GC-rich polymerase mix, and 1.9 μL FAM-Primers were added to each reaction and PCR was performed with the following cycles: 1 cycle of 95 °C for 5 minutes, 27 cycles involving 97 °C for 35 seconds, 62 °C for 35 seconds, and 72 °C for 4 minutes, 1 cycle of 72 °C for 10 minutes followed by cooling to 4 °C. 10 μL of each PCR reaction was then mixed with 2 μL Gel Loading Dye, Purple (6X) (NEB) and run at 80V for 20 minutes and 100V for 60 minutes on a 1% agarose gel followed by staining with ethidium bromide solution for 15 minutes and visualization with Molecular Imager® Gel Doc™ XR+ (Bio-Rad). Quantification of band intensities was achieved with the Gel Analysis utility in ImageJ software.

Viral integration site detection

Viral integration sites were defined by following a pipeline developed by Ho et al. [461] with several key modifications. First, quality trimmed WGBS reads (from above) were aligned with bowtie2 v2.3.4.1 (--very-sensitive-local option) to custom FASTA files containing *in silico* bisulfite-converted sequences (CG to YG, C to T) of forward and reverse strands of the integration-capable lentiviral elements (between two LTRs) from all treatments for that particular cell line: dCas9 plasmids, gRNA3 or gRNAscr plasmids,

and Cre plasmids. Notably, the sequence from the lentiviral dCas9 plasmid sequence was *in silico* recombined (deletion between loxP sites, leaving one loxP site) to mimic Cre action in the cells. Then samtools v1.3.1[462] was invoked to extract all aligned soft-clipped reads; these are reads that were clipped in order to align to the lentiviral sequences and therefore the clipped portion represents possible read-through into mouse genome (no difference from Ho et al.). We then ran a modified variant of the script published by Ho et al. (to allow for alignment to mouse bisulfite converted genomic sequences generated by Bismark rather than human unconverted genomic sequences) that used BLAST[463] to identify boundaries between viral and mouse sequences (Supplementary Software 1). All overlaps with dmCpGs were performed with BEDTools intersect v2.29.2[464].

Statistics and data visualization

All data involving simple statistical tests not described above in WGBS and ChIP-seq methodology (e.g. T-test, Mann-Whitney test, Pearson's r, Holm-Sidak correction for multiple testing) were calculated and graphed with Graphpad Prism 8 software.

12.6 Data availability

The whole genome bisulfite sequencing (WGBS) data generated in this study have been deposited in the Gene Expression Omnibus (GEO) database under accession code GSE162138 (https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE162138). The chromatin immunoprecipitation sequencing (ChIP-seq) data generated in this study have been deposited in the GEO database under accession code GSE174275 (https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE174275). Source data are provided with the online open access version of this article. Mouse mm10 genome is available at https://www.ncbi.nlm.nih.gov/assembly/GCF_000001635.20/. Publicly available microarray data used for candidate gene selection for **Supplementary Table**

7 is in the GEO database under accession code GSE8374 (https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE8374).

12.7 Code availability

Code used to find viral integration sites is available as Supplementary Software 1, available with the published online version of this article.

12.8 Acknowledgements

This study was funded by the Canadian Institutes of Health Research (PJT159583). D.M.S was supported by fellowships from the McGill University Faculty of Medicine (Friends of McGill Fellowship; JP Collip Fellowship in Medical Research; James Frosst Fellowship). We thank the Next Generation Sequencing team of the Centre d'expertise et de services Génome Québec at McGill University for their sequencing services. We also thank Julien Leconte for performing cell sorting at the McGill University Flow Cytometry Core Facility and Andrew Bayne of McGill University for assistance in visualization of the CRISPR/Cas9 structure.

12.9 Author Contributions

D.M.S. and M.S. designed all experiments. D.M.S. performed all experiments and created all of the figures. D.M.S. and M.S. both contributed to data analysis and writing of the manuscript.

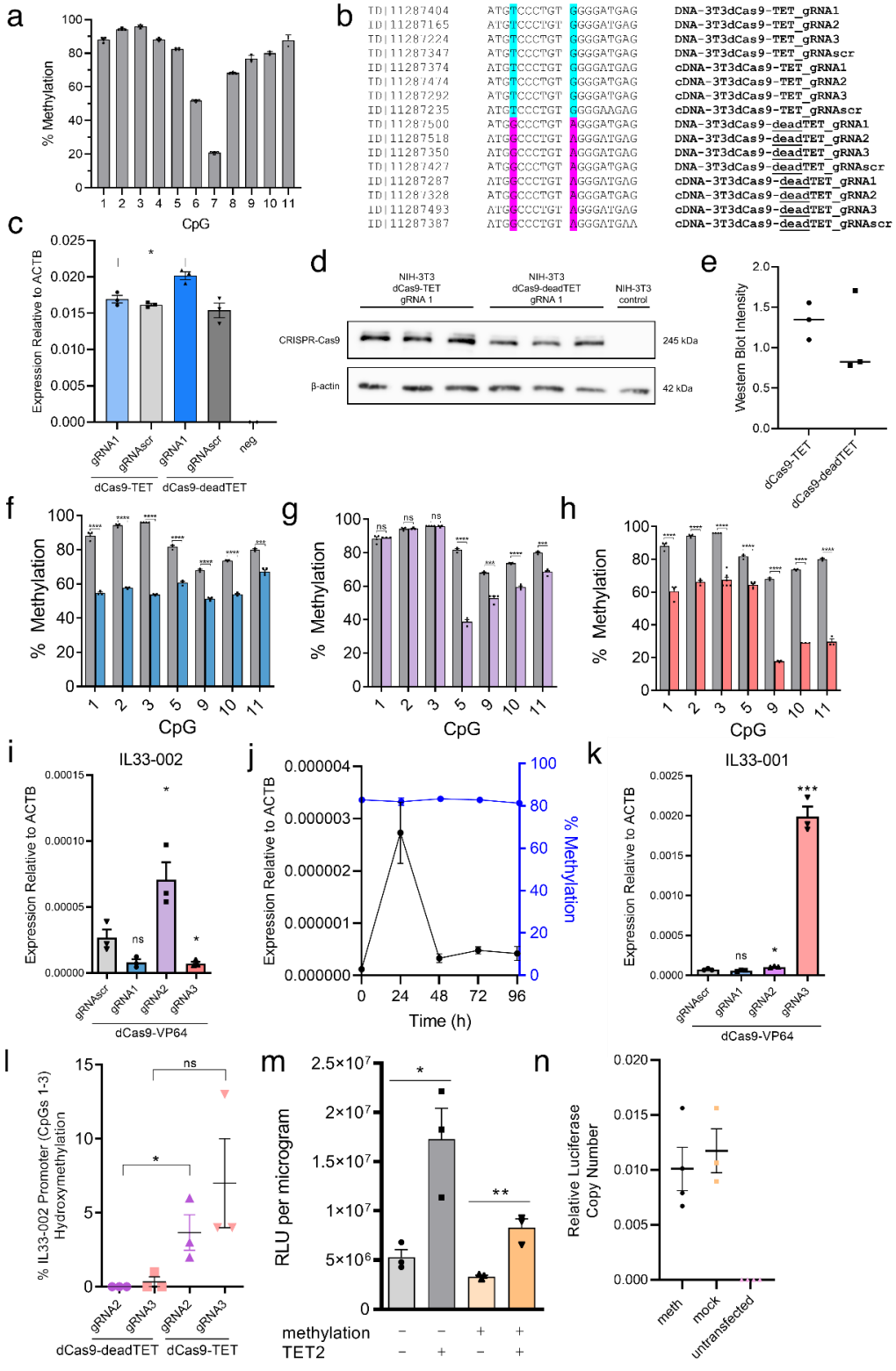
12.10 Competing interests

D.M.S. and M.S. have no competing interests to declare.

12.11 Supplementary Information

This supplementary information contains Supplementary Figures 1-11 and Supplementary Tables 1-7. Supplementary Data 1-4, Supplementary Software 1, and Source Data can be accessed via following link: https://www.nature.com/articles/s41467-021-25991-9

Supplementary Figure 1

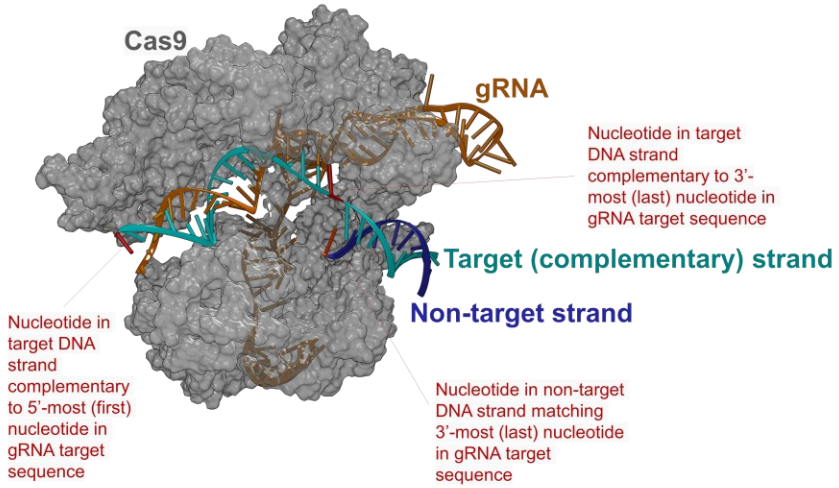


Supplementary Figure 1. Confounds involved in CRISPR/TET-based approaches. (A) Percent of DNA methylation (mean ± SEM) assayed by bisulfite-pyrosequencing at 11 CpGs in the *Il33-002* promoter of control, untreated NIH-3T3 cells (n = 3 independent samples for CpGs 1-8, n = 4 independent samples for CpGs 9-11). (B) Aligned Sanger sequencing results of region of TET bearing the inactivating mutation in deadTET controls from one representative cell line from triplicate treatments in Figure 1 with Sanger ID (left), DNA sequence (middle), and source of DNA (right). (C) mRNA expression (mean ± SEM) of dCas9-TET or dCas9-deadTET (single primer pair that amplifies common region) relative to Actb in corresponding NIH-3T3 cells from Figure1E-I expressing either dCas9-TET or dCas9-deadTET in combination with either gRNAscr or gRNA1. Negative control cells are untreated NIH-3T3 (n = 3 independent samples, n = 2 independent samples for negative control). (D) Western blot with anti-CRISPPR/Cas9 (top panel) or anti-β-actin (bottom panel) antibody in NIH-3T3 cells expressing dCas9-TET and gRNA1 (n = 3 independent experiments), dCas9-deadTET and gRNA1 (n = 3 independent experiments), or negative control (untreated) NIH-3T3 cells (n = 1). (E) Quantification of (D) using ImageJ involving normalization of anti-CRISPR/Cas9 antibody signals to anti-β-actin antibody signals (mean, n = 3 independent experiments) (F-H) Percent of DNA methylation assayed by bisulfite-pyrosequencing at 7 targeted CpGs in NIH-3T3 cells treated with dCas9-VP64 and either gRNA1 (F; blue, n = 3-5 independent experiments), gRNA2 (G; purple, n = 3-4 independent experiments), gRNA3 H; pink, n = 3-6 independent experiments) or gRNAscr (grey; identical data in B-D, shown for comparison) (mean ± SEM) (n varies depending on specific condition and CpG; see Source Data file for specific n of interest). (I) Expression of *Il33-002* (mean ± SEM) quantified by RT-qPCR and normalized to *Actb* expression in NIH-3T3 stably expressing one of 4 gRNAs and dCas9-VP64. Statistical comparisons are to gRNAscr condition (n = 3 independent experiments). (J) (Left y-axis) *Il33-002* gene expression (mean ± SEM) quantified by RT-qPCR and normalized to *Actb* expression in NIH-3T3 cells stably expressing dCas9-VP64 and transiently transfected with gRNA2 and harvested after the indicated number of hours.

(Right y-axis) DNA methylation level (mean ± SEM) of CpG 11 at the TSS of the *Il33-002* promoter in the corresponding cells (n = 3 independent samples). (K) Expression of *Il33-001* (mean ± SEM) quantified by RT-qPCR and normalized to *Actb* expression in NIH-3T3 stably expressing one of 4 gRNAs and dCas9-VP64. Statistical comparisons are to gRNAscr condition (n = 3 independent experiments). (L) Percent of DNA hydroxymethylation (mean ± SEM) assayed by KRuO4 oxidation of DNA followed by bisulfite-pyrosequencing in parallel with unoxidized controls and calculated as decrease in methylation after oxidation at CpGs 1, 2, and 3, averaged, which are distant from the gRNA2 (purple) and gRNA3 (pink) binding sites, under the specified stable treatments in NIH-3T3 cells (x-axis) (n = 3 independent experiments). (M) Relative light units normalized to extracted total protein quantity (mean ± SEM) in transfected HEK293 cells. Cells were transiently transfected with methylated or unmethylated SV40-luciferase vector along with mammalian TET2 expression plasmid or empty vector (pcDNA3.1) control (n = 3 independent experiments) (N) Cellular luciferase gene (DNA) copy number (mean ± SEM), measured by qPCR with primers in Supplementary Table S2, and normalized to levels of genomic *Actb*, for untransfected cells and transfections of 50 ng of SV40-pCpGl, either mock methylated or fully methylated by M.SssI (n = 3 independent experiments). * indicates statistically significant difference of $P < 0.05$, ** of $P < 0.01$, *** of $P < 0.001$, **** of $P < 0.0001$, and ns = not significant (Student's t-test, two-sided, with Holm-Sidak correction if number of tests is greater than 3).

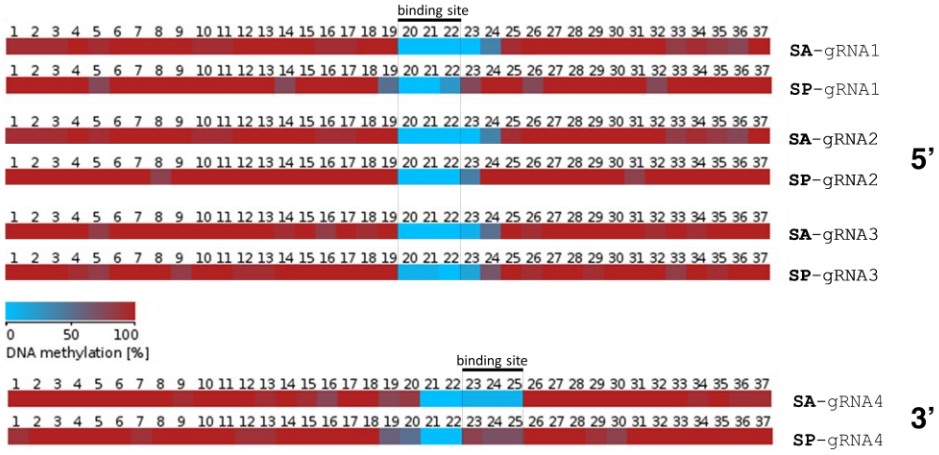
Supplementary Figure 2

a

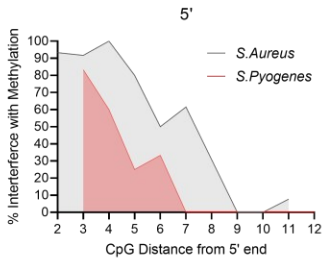


b

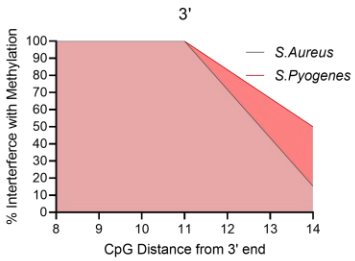
c



d

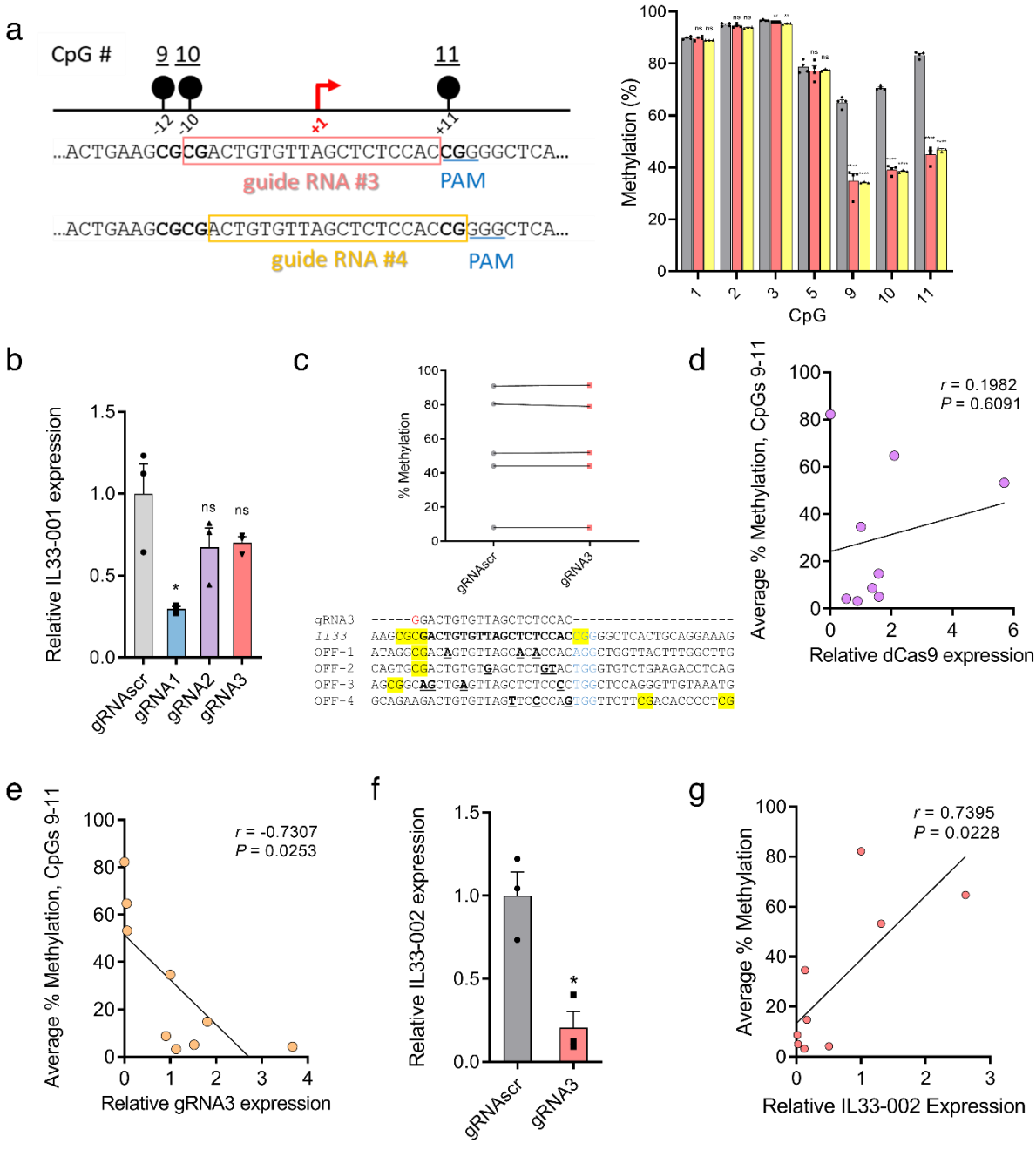


e



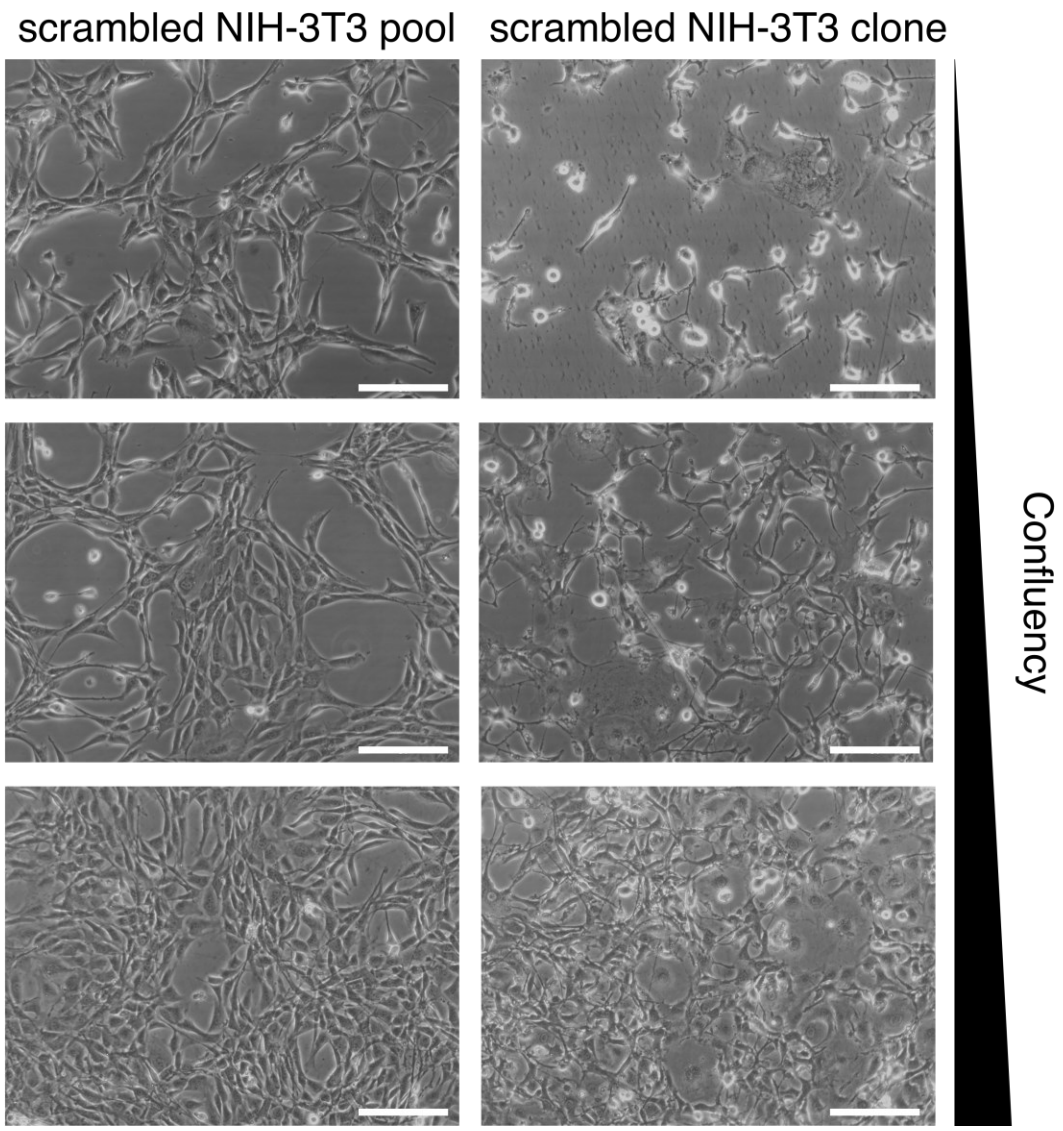
Supplementary Figure 2. Differences in steric interference with DNA methyltransferases by Cas orthologs. (A) Cryo-EM structure of Cas9 (grey) in complex with gRNA (orange), target DNA strand (light blue), and non-target DNA strands (dark blue) (partial) from [42] (PDB: 6O0Z). Two 5' prime nucleotides (CC in target DNA strand, GG in gRNA) that were not resolved in the cryo-EM structure were built and energy minimized using Molecular Operating Environment (MOE) software. The final figure was generated in UCSF Chimera. Red nucleotides are labeled accordingly and represent the nucleotides in the DNA complementary to either the first (5') nucleotide of the gRNA target sequence or the last (3') nucleotide of the gRNA target sequence as well as its complement in the non-target strand. (B) Diagram and sequence of *CDKN2A* region targeted by additional gRNAs. Due to the display of the reverse complement sequence of Figure 3, CpGs are numbered differently (black) but Figure 3 numbering system is shown below in red. gRNA sequences are shown in green, where the entire green sequence represents the S. pyogenes gRNA and the addition 5' nucleotide in black represents the additional nucleotide needed for S. aureus gRNAs. S. aureus PAM site is shown in blue, with the first 3 nucleotides (5' to 3') represent the NGG PAM of the S. pyogenes gRNA. (C) Each horizontal row depicts a heatmap of average DNA methylation at each numbered CpG over 10-20 (except SP-gRNA4, 4 clones) individual strands of DNA (bisulfite-converted clones) where light blue represents 0% methylation and dark red represents 100% methylation. The CpGs within the binding site of the labeled gRNA are labeled and enclosed in dashed lines. gRNAs1-3 interrogate DNA methylation interference of 5' proximal CpGs and gRNA4 interrogates that of 3' proximal CpGs. Lowly methylated strands of DNA (poor M.SssI methylation) and strands with unaffected binding sites (unbound by dCas9) were excluded from the analysis because efficacy was not under evaluation. (D-E) Data from (C) transformed into a percent methylation as a function of CpG distance in base pairs from the 5' (D) or 3' (E) end of the gRNA sequence (including PAM) and S. aureus(grey) or S. pyogenes (pink) across gRNAs 1-3 (D) or gRNA4 (E).

Supplementary Figure 3



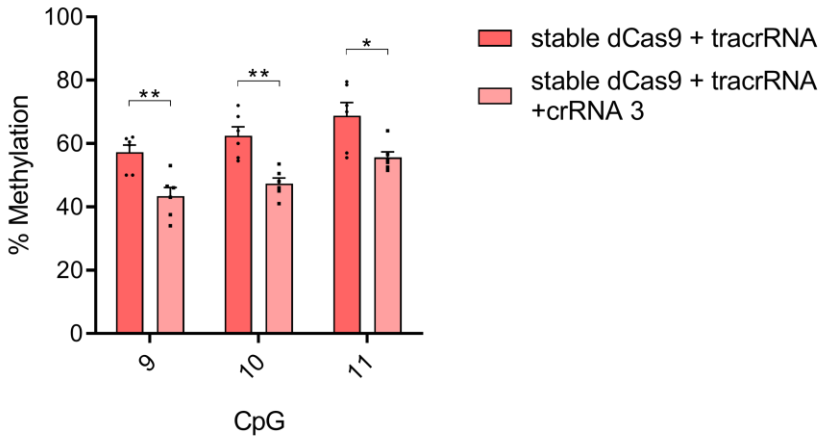
Supplementary Figure 3. Characteristics of dCas9-based inhibition of methylation at the Il33 locus. (A) (Left) Diagram of *Il33-002* promoter with location of gRNA3 and gRNA4. TSS is marked by a red arrow and CpGs are marked by black circles. In sequence below, PAM represents protospacer adjacent motif, gRNA sequences are boxed, and CpGs are bolded. (Right) Methylation levels (mean ± SEM) assessed by pyrosequencing of NIH-3T3 cells expressing dCas9 and gRNA3 (pink) or gRNA4 (yellow). Values displayed as mean +/- SEM (n = 3 independent experiments). (B) Il33-001 expression (mean ± SEM) in NIH-3T3 cell lines stably expressing gRNAscr or one of 3 Il33-002-targeting gRNAs in combination with dCas9, assayed by qRT-PCR and normalized to *Actb* expression (n = 3 biologically independent samples). (C) Comparison of methylation levels, assayed by pyrosequencing, of 5 top off-target CpGs in NIH-3T3 cell lines stably expressing scrambled gRNA or gRNA 3 in combination with dCas9. (D) Correlation of 9 dCas9:gRNA3 clones from (Fig 4E); x-axis displays dCas9 expression normalized to *Actb* expression; y-axis displays average methylation at CpGs 9, 10, and 11 in each clone assayed by pyrosequencing ($r = 0.1982$, $P = 0.6091$). (E) Correlation of 9 dCas9:gRNA3 clones from (Fig 4E); x-axis displays gRNA3 expression normalized to *Actb* expression; y-axis displays average methylation at CpGs 9, 10, and 11 in each clone assayed by pyrosequencing ($r = -0.7307$, $P < 0.05$). (F) Il33-002 expression (mean ± SEM) in NIH-3T3 cell lines stably expressing scrambled gRNA or gRNA3 in combination with dCas9, assayed by qRT-PCR and normalized to Actb expression (n = 3 biologically independent experiments, * indicates $P < 0.05$ vs gRNAscr, t-test). (G) Scatter plot of 9 dCas9:gRNA3 clones from (4E); x-axis displays relative Il33-002 expression as assayed in (F); y-axis displays average methylation at CpGs 9, 10, and 11 in each clone assayed by pyrosequencing ($r = 0.7395$, $P = 0.0228$). * indicates statistically significant difference of $P < 0.05$, ** of $P < 0.01$, *** of $P < 0.001$, **** of $P < 0.0001$, and ns = not significant (Student's t-test, two-sided, with Holm-Sidak correction if number of tests is greater than 3).

Supplementary Figure 4



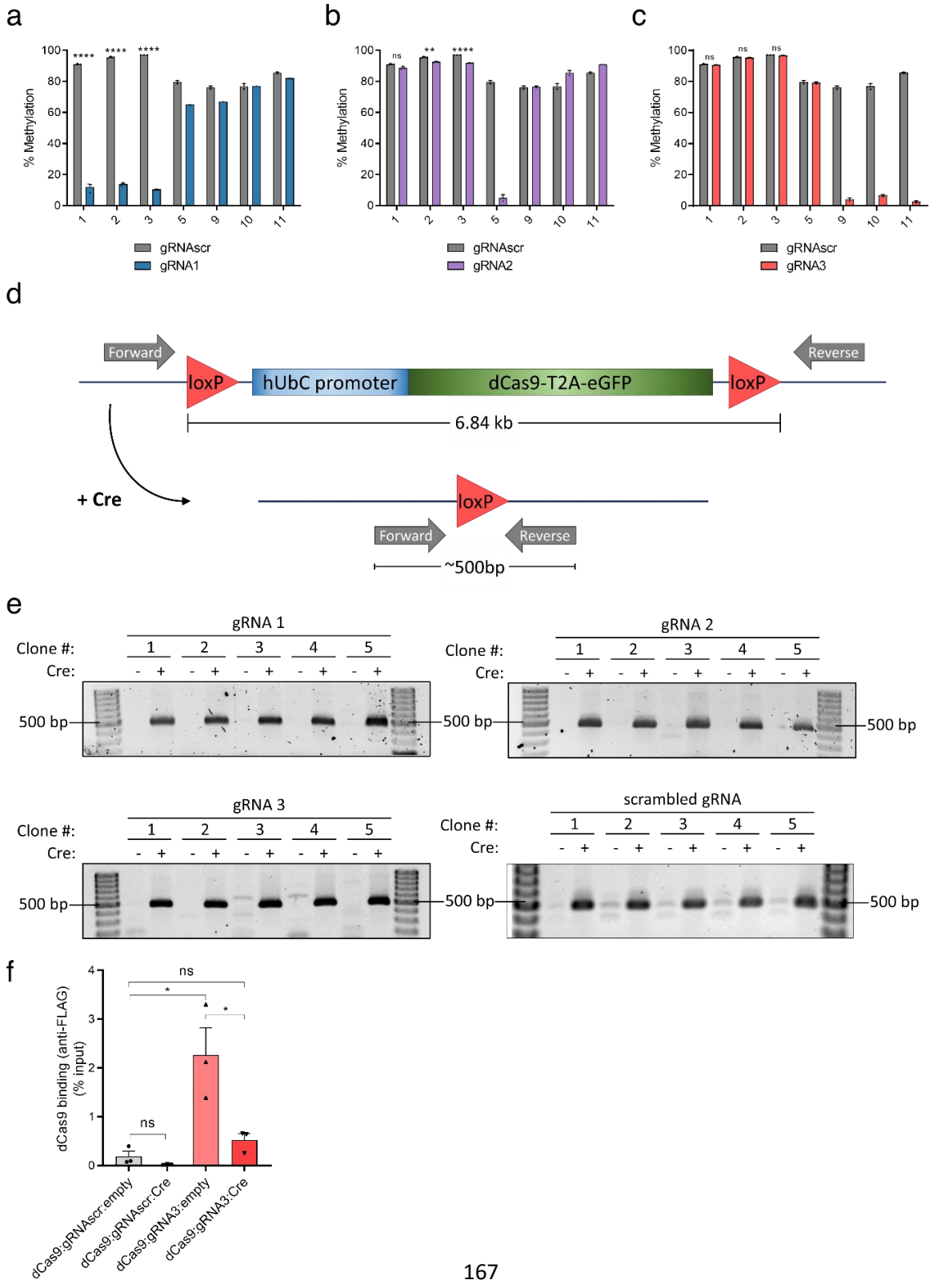
Supplementary Figure 4. Clonal selection is a deficient method for derivation of cell lines with effective dCas9-based demethylation. Light microscopy images demonstrating the appearance of normal healthy NIH-3T3 pools (3 left panels) with increasing confluency downwards. In comparison, morphological irregularities can be seen after clonal isolation from the same source cells in 3 distinct clonal populations at 3 different levels of confluency (n = 1 independent clonal cell lines for each level of confluency for a total of 3 independent cell lines per experimental condition). Scale bars represent 100 µM.

Supplementary Figure 5



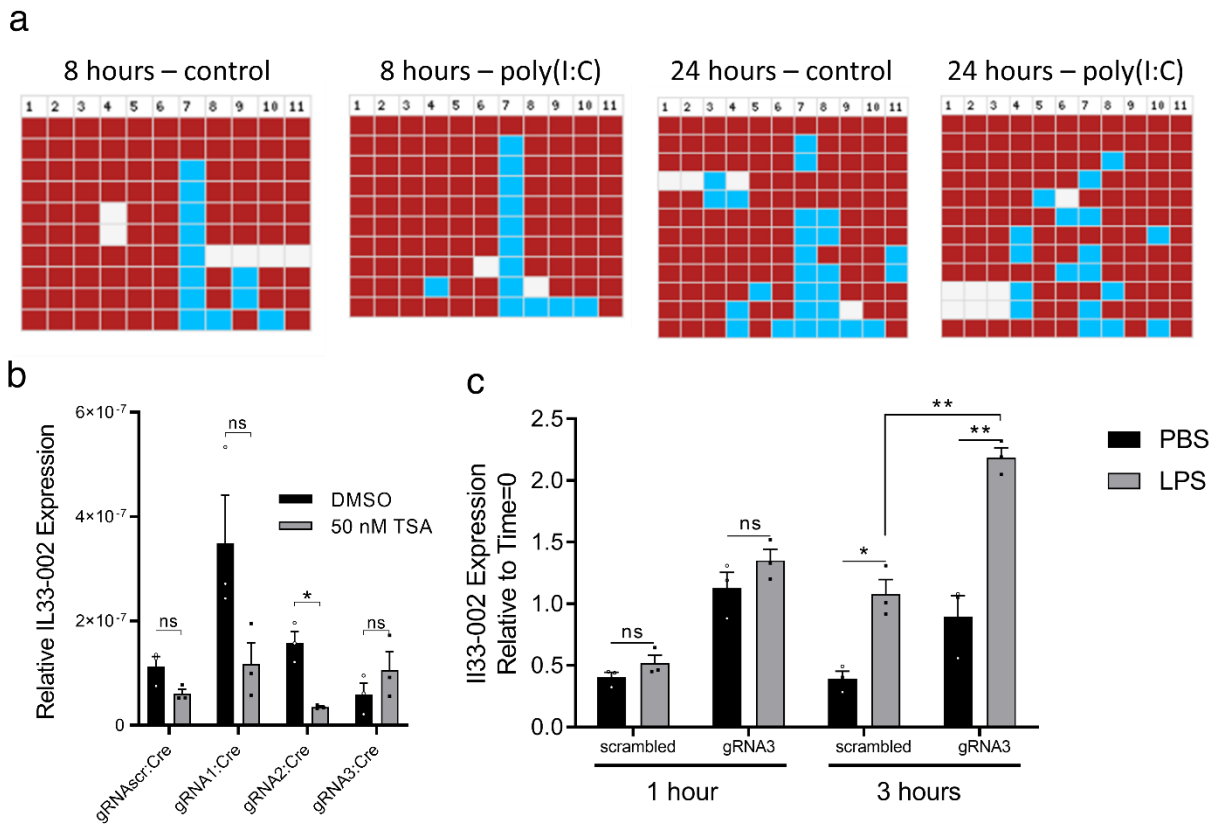
Supplementary Figure 5. Transient transfection of gRNA components. Methylation levels assessed by bisulfite-pyrosequencing (mean ± SEM) of target CpGs 9, 10, and 11 after NIH-3T3 cells stably expressing dCas9 were transiently transfected (Xtremegene siRNA transfection reagent, Sigma) with either tracrRNA alone (red) or tracrRNA and crRNA3 (pink), a two-component version of gRNA3 (n = 6 biologically independent experiments). * indicates statistically significant difference of $P < 0.05$ and ** of $P < 0.01$(Student's t-test, two-sided, with Holm-Sidak correction).

Supplementary Figure 6



Supplementary Figure 6. Verification of success of removable lentiviral dCas9 strategy. (A-C) DNA methylation levels (mean ± SEM) by pyrosequencing of NIH-3T3 cells stably expressing gRNA1 (A), gRNA2 (B), or gRNA3 (C) and floxed dCas9 in order to validate targeted demethylation (n = 1-3 biologically independent samples). (D) Diagram of the dCas9 expression construct. It is flanked by loxP sites that facilitate recombination and deletion by Cre recombinase. Forward and reverse PCR primers lie outside the loxP sites such that a 6.84 kb product could be made when Cre is not present in the cells and if PCR extension times are increased to allow this product to form. After removal of the dCas9 expression cassette by Cre recombinase, the same PCR primers create a product of approximately 500 base pairs in size. (E) Agarose gels showing recombination-dependent PCR products using the primers in (D) in n = 5 independent cell lines stably expressing dCas9 and each indicated gRNA after independent treatment by empty virus (-) or Cre recombinase (+). A 500 base pair product is visible in each Cre-containing lane. Primers are listed in Supplementary Table S2. (F) Chromatin immunoprecipitation of cells from Figure S6E with antibody against the 5' 3XFlag-tagged dCas9 (anti-Flag antibody) followed by qPCR using primers surrounding the Il33-002 TSS. dCas9 binding is expressed as percent input (n = 3 biologically independent experiments). * indicates statistically significant difference of $P < 0.05$, ** of $P < 0.01$, *** of $P < 0.001$, **** of $P < 0.0001$, and ns = not significant (Student's t-test, two-sided, with Holm-Sidak correction if number of tests is greater than 3).

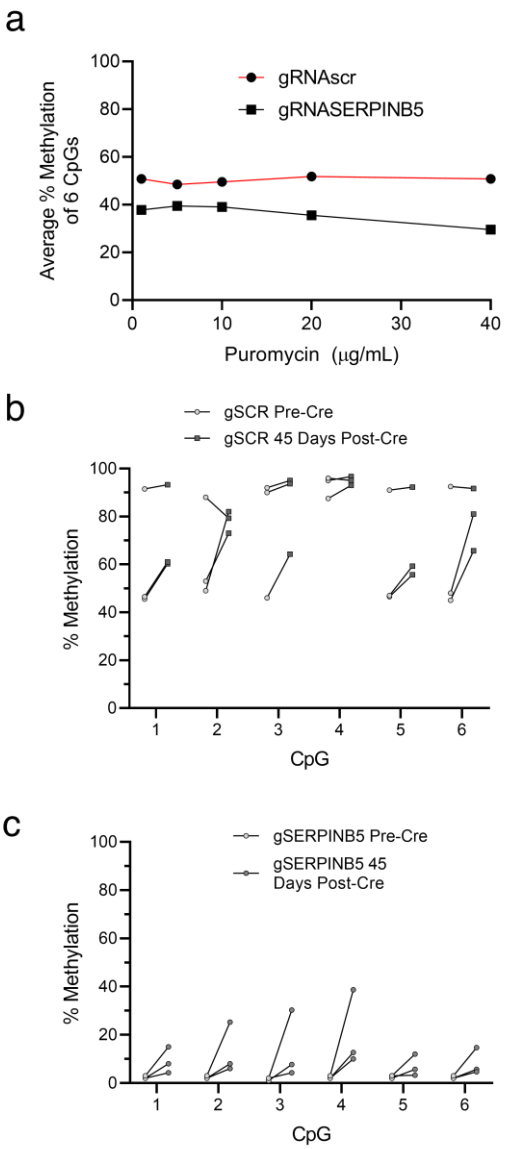
Supplementary Figure 7



Supplementary Figure 7. Effect of inducing agents on *Il33*. (A) Bisulfite-cloning and sanger sequencing analysis of the *Il33-002* promoter in NIH-3T3 cells treated with 1 µg/mL poly(I:C) or water control for 8 or 24 hours. Each horizontal row is one strand of DNA. Numbers indicate the CpG in the promoter. Red squares indicate methylated CpGs, blue squares indicate unmethylated CpGs, and white squares indicate a lack of data due to sequencing failure. (B) *Il33-002* expression (mean ± SEM) in 50nM TSA or vehicle (DMSO) treated NIH-3T3 cell lines stably expressing gRNAscr, gRNA1, gRNA2, or gRNA3 under high-puromycin conditions in combination with dCas9, followed by dCas9 removal by Cre recombinase as assayed by qRT-PCR and normalized to *Actb* expression (n = 3 biologically independent experiments). (C) *Il33-002* expression (mean ± SEM) in 100 ng/mL lipopolysaccharide (LPS) or vehicle (PBS) treated NIH-3T3 cell lines stably expressing gRNAscr or gRNA3 under high-puromycin conditions in

combination with dCas9, followed by dCas9 removal by Cre recombinase, as assayed by qRT-PCR and normalized to *Actb* expression, either 1 or 3 hours after treatment and displayed relative to expression measured at time=0 (n = 3 biologically independent experiments). * indicates statistically significant difference of $P < 0.05$, ** of $P < 0.01$, *** of $P < 0.001$, **** of $P < 0.0001$, and ns = not significant (Student's t-test, two-sided, with Holm-Sidak correction if number of tests is greater than 3).

Supplementary Figure 8

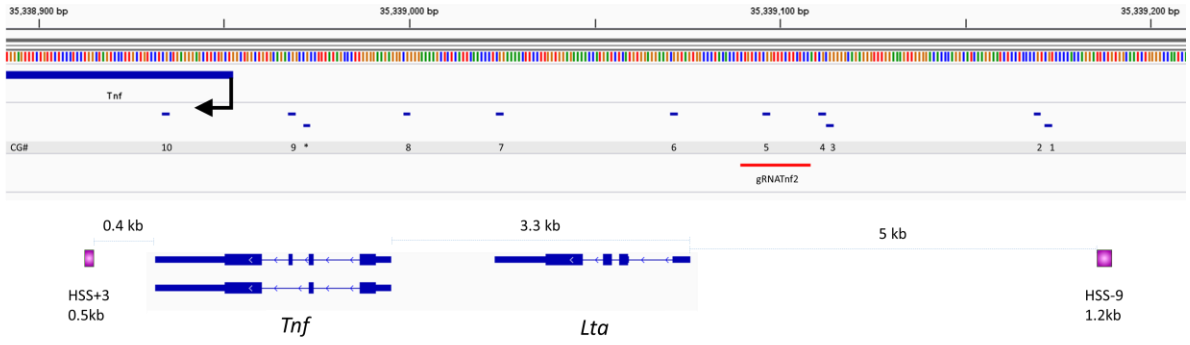


Supplementary Figure 8. Methylation levels of *SERPINB5*. (A) DNA methylation levels (mean ± SEM) assessed by bisulfite-pyrosequencing of CpGs 1-6 in the *SERPINB5* promoter in MDA-MB-231 cell lines stably expressing dCas9 and either gRNAscr (red) or gRNASERPINB5 (black), averaged across all 6 CpGs and plotted as a function of increasing puromycin concentration (n = 1 per puromycin concentration). (B-C) Percent DNA methylation assessed by bisulfite-pyrosequencing of n = 3 MDA-MB-231 clonal cell lines expressing dCas9 and gSCR (B) or gRNASERPINB5 (C) prior

to dCas9 excision by Cre (light grey) and 45 days after the end of selection for Cre recombinase (dark grey) (mean ± SEM).

Supplementary Figure 9

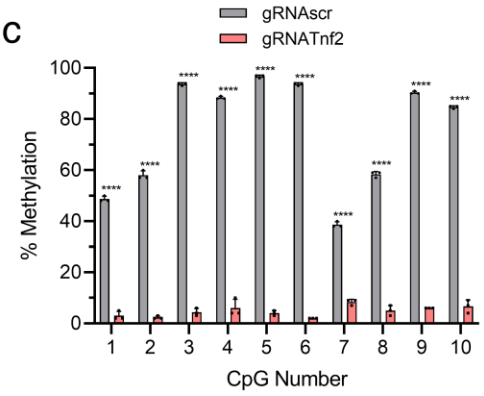
a



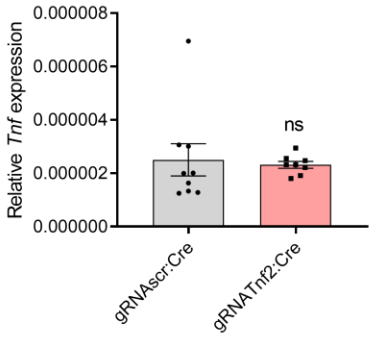
b

gRNA	gRNAscr	1	2	3	4	5	6
CpG target		1,2	3 to 5	6	7	8,10	11
CpG							
1	49	3.33	2.7	1.7	43.7	51	30
2	58	4.33	3	2.3	48.3	60	28.3
3	93.7	58	4.3	93.7	79.7	93	90.3
4	88.3	81	6	86.7	72.7	88.3	81.3
5	96.7	65.3	4	92	87.3	88.7	94.3
6	93.7	64.7	7.7	28	93.6	81.7	83
7	38.7	31.7	2	37.7	23.7	51.7	40.7
8	58.3	92	6	11	52.3	52	52.3
9	90.3	83.7	5	22	85.6	77.3	91.3
10	84.7	68	6.7	71.3	60	83.7	66

c

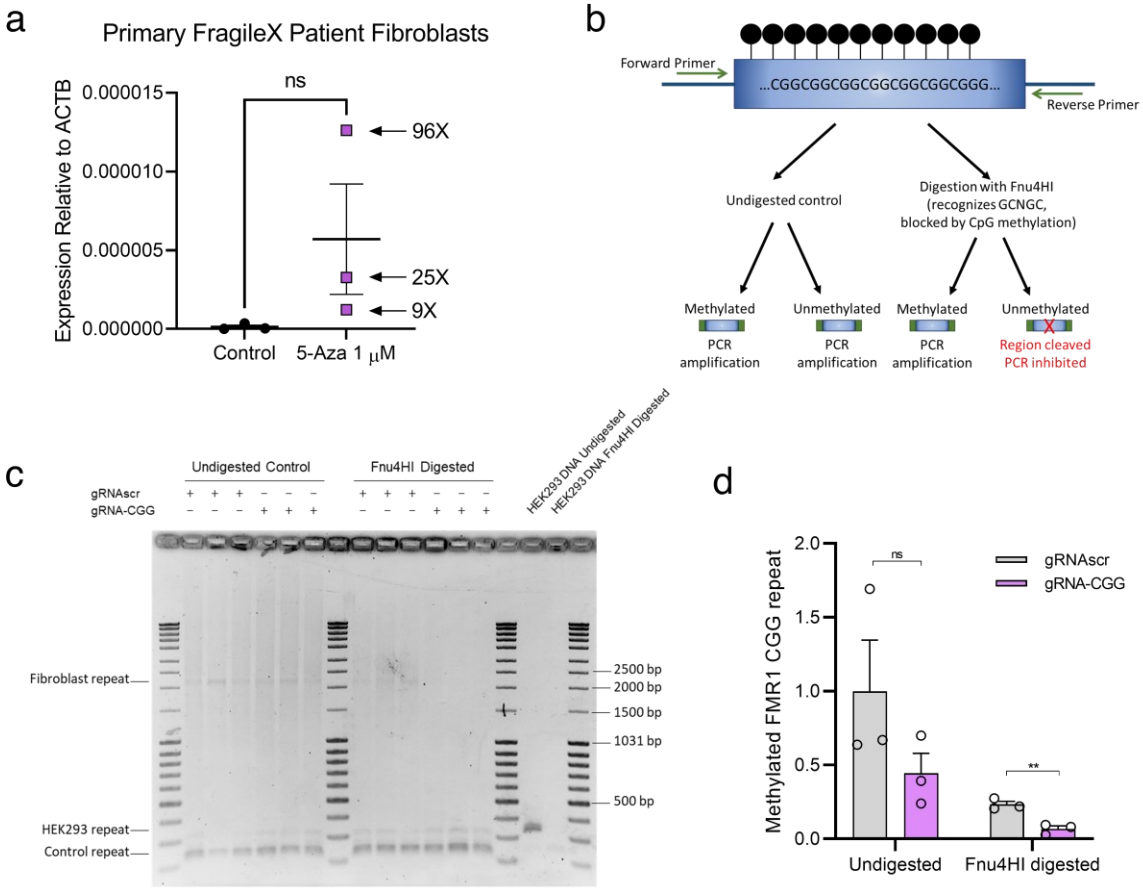


d



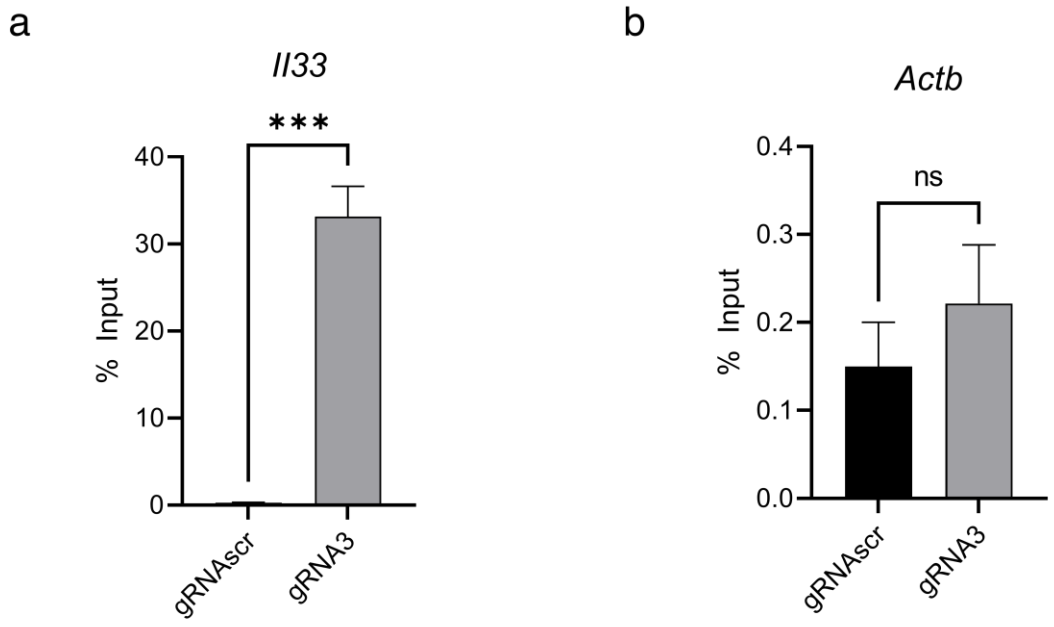
Supplementary Figure 9. Demethylation of the *Tnf* promoter. (A) Genome browser view of the murine Tnf locus;(Top) each CG location marked by a blue dash and numbered below, TSS indicated by a black arrow., and the location of gRNATnf2 is labeled with a red line and marked accordingly; (Bottom) Two known distal enhancers of *Tnf* expression indicated with purple boxes, named and marked with distances to Tnf TSS [77]. (B) Table demonstrating the average methylation of *Tnf* CpGs numbered in (A) as measured by bisulfite-pyrosequencing a function of six candidate Tnf-targeting gRNAs or gRNAscr control in NIH-3T3 cells also stably expressing dCas9. CpGs within the gRNA binding site are indicated below the gRNA number and their methylation status is highlighted in yellow in the corresponding gRNAs. (C) DNA methylation levels (mean ± SEM) assessed by bisulfite-pyrosequencing of NIH-3T3 cells stably expressing dCas9 and either gRNATnf2 (pink) or gRNAscr (grey) (n = 3 biologically independent experiments). (D) Tnf expression (mean ± SEM) in NIH-3T3 cell lines subcloned from those in Figure 7H stably expressing gRNAscr (grey) or gRNATnf2 (pink) under high-puromycin conditions in combination with dCas9, followed by dCas9 removal by Cre recombinase as assayed by RT-qPCR and normalized to *Actb* expression (n = 9 independent clones for gRNAscr, n = 8 independent clones for gRNATnf2). * indicates statistically significant difference of $P < 0.05$, ** of $P < 0.01$, *** of $P < 0.001$, **** of $P < 0.0001$, and ns = not significant (Student's t-test, two-sided, with Holm-Sidak correction if number of tests is greater than 3).

Supplementary Figure 10



Supplementary Figure 10. Demethylation of FMR1 promoter. (A) Expression of the *FMR1* gene (mean ± SEM) relative to GAPDH expression in primary fibroblasts from Fragile X patient, measured by RT-qPCR after control (water) or 5-aza-2'-deoxycitidine treatment. The difference is not statistically significant (n = 3 biologically independent experiments, P = 0.1, Mann-Whitney test, two-sided). (B) Schematic of the experimental workflow used to determine methylation status of the *FMR1* CGG repeat region. DNA from Fragile X patient fibroblasts is first either subjected to digestion with the methylation sensitive restriction enzyme Fnu4HI or a control (no enzyme) reaction, followed by DNA cleanup and amplification by primers that sit immediately 5' and 3' of the CGG repeat region. A reduction in methylation can be observed by increased DNA digestion by Fnu4HI and failure to amplify by PCR as visualized by agarose gel electrophoresis. (C) Agarose gel electrophoresis results after PCR for the workflow in (B). Lanes 1, 8, 15, and 18 contain MassRulerTM DNA Ladder Mix (Thermo Fisher) and select band sizes are labeled to the right of the gel. Samples are DNA extracted from Fragile X patient fibroblasts expressing dCas9 and either gRNA-CGG or gRNAscr as indicated above the gel. DNA is either undigested control (left 6 samples) or Fnu4HI digested (right 6 samples). Undigested HEK293 control DNA with a shorter repeat region and Fnu4HI digested HEK293 control DNA after PCR are in lanes 16 and 17, respectively. Amplicon size in Fragile X patient fibroblasts is labeled to the left of the gel between 2000 and 2500 bp. Amplicon size in control HEK293 cells is between 300 and 400 bp and is labeled to the left of the gel. Spike-in control DNA to ensure successful PCR is also labeled to the left. (D) Quantification of results in (C) achieved by normalization of intensity of genomic fibroblast amplicon to that of control amplicon using ImageJ software; data is expressed as intensity relative to the undigested dCas9:gRNAscr condition (mean ± SEM). ** indicates statistical significance at P < 0.01 and ns indicates no statistically significant difference by Student's t-test,two-sided (n = 3 biological replicates).

Supplementary Figure 11



Supplementary Figure 11. Validation of ChIP with anti-FLAG antibody prior to ChIP-seq. (A) Pulldown of Il33-002 locus measured by qPCR of an amplicon near the transcription start site following ChIP with anti-FLAG antibody in n = 3 independent NIH-3T3 cell lines expressing FLAG-tagged dCas9 and either gRNAscr or gRNA3. Data is expressed as percent input (mean ± SEM) (B). Same as (A) except primers were used for the *Actb* locus, which should not be bound by FLAG-tagged dCas9 in either of the two treatment groups. *** indicates $P < 0.001$ and ns indicates no statistical significance using the two-sided Student's t-test.

Il33 gRNA 1	GAGCCGGTGTTTTCTTGAGC
Il33 gRNA 2	GGTGTGACATAGCCCCATAG
Il33 gRNA 3	GGACTGTGTTAGCTCTCCAC
Il33 gRNA 4	GCTGTGTTAGCTCTCCACCG
Il33 gRNA 5	GCACTCACCTCAATACAGAC
Il33 gRNA 6	GAGCTGATAGATGCTACTAT
scrambled gRNA	GCACTACCAGAGCTAACTCA
HNF4A gRNA	GGGCGCGTTCACGCTGACCA
SERPINB5 gRNA	GAGGAGTGCCGCCGAGGCG
p16/CDKN2A gRNA (CpG 17)	GCATGGAGCCTTCGGCTGAC
SA-gRNA1 (CDKN2A)	AGCAGCATGGAGCCGGCGGCG
SA-gRNA2 (CDKN2A)	GAGCAGCATGGAGCCGGCGGC
SA-gRNA3 (CDKN2A)	GGAGCAGCATGGAGCCGGCGG
SA-gRNA4 (CDKN2A)	AGCGGGCGGCGGGGAGCAGCA
SP-gRNA1 (CDKN2A)	GCAGCATGGAGCCGGCGGCG
SP-gRNA2 (CDKN2A)	AGCAGCATGGAGCCGGCGGC
SP-gRNA3 (CDKN2A)	GAGCAGCATGGAGCCGGCGG
SP-gRNA4 (CDKN2A)	GCGGGCGGCGGGGAGCAGCA
gRNA-CGG (FMR1)	GGCGGCGGCGGCGGCGGCGG
gRNATnf2	GGAGAAGAAACCGAGACAG
gBlock sequence (20bp poly-N is replaced with gRNA sequence)	TGTACAAAAAAGCAGGCTTTAAAGGAACCAA TTCAGTCGACTGGATCCGGTACCAAGGTCG GGCAGGAAGAGGGCCTATTTCCCATGATTC CTTCATATTTGCATATACGATACAAGGCTGTT AGAGAGATAATTAGAATTAATTTGACTGTAAA CACAAAGATATTAGTACAAAATACGTGACGT AGAAAGTAATAATTTCTTGGGTAGTTTGCAG TTTTAAAATTATGTTTTAAAATGGACTATCAT ATGCTTACCGTAACTTGAAAGTATTTCGATTT

	CTTGGCTTTATATATCTTGTGGAAAGGACGA AACACCNNNNNNNNNNNNNNNNNNNGTTT TAGAGCTAGAAATAGCAAGTTAAAATAAGGC TAGTCCGTTATCAACTTGAAAAAGTGGCACC GAGTCGGTGCTTTTTTTCTAGACCCAGCTTT CTTGTACAAAGTTGGCATTA

Supplementary Table 1. Target sequences of all gRNAs used in the study.

Application	**Primer Name**	**Primer Sequence**
In vitro gRNA synthesis	Il33gRNA1_IVT_F	TAATACGACTCACTATAGAGCCGGTGT TTTCTTGAGC
	Il33gRNA1_IVT_R	TTCTAGCTCTAAAACGCTCAAGAAAAC ACCGGCT
	Il33gRNA2_IVT_F	TAATACGACTCACTATAGGTGTGACAT AGCCCCATAG
	Il33gRNA2_IVT_R	TTCTAGCTCTAAAACCTATGGGGCTAT GTCACAC
	Il33gRNA3_IVT_F	TAATACGACTCACTATAGGACTGTGTT AGCTCTCCAC
	Il33gRNA3_IVT_R	TTCTAGCTCTAAAACGTGGAGAGCTAA CACAGTC
	Il33gRNA5_IVT_F	TAATACGACTCACTATAGCACTCACCT CAATACAGAC
	Il33gRNA5_IVT_R	TTCTAGCTCTAAAACGTCTGTATTGAG GTGAGTG

	Il33gRNA6_IVT_F	TAATACGACTCACTATAGAGCTGATAG ATGCTACTAT
	Il33gRNA6_IVT_R	TTCTAGCTCTAAAACATAGTAGCATCTA TCAGCT
	gRNAscrambled_IV T_F	TAATACGACTCACTATAGCACTACCAG AGCTAACTCA
	gRNAscrambled_IV T_R	TTCTAGCTCTAAAACTGAGTTAGCTCT GGTAGTG
	p16_IVT_gRNA1F	TAATACGACTCACTATAGCATGGAGCC TTCGGCTGAC
	p16_IVT_gRNA1R	TTCTAGCTCTAAAACGTCAGCCGAAGG CTCCATGC
gBlock amplification and A-tailing	gBlockgRNA_F	TGTACAAAAAAGCAGGCTTTAAAG
	gBlockgRNA_R	TAATGCCAACTTTGTACAAGAAAG
Cre Recombination – Removal of dCas9	dCas9_recomb_F	ATCGTTTCAGACCCACCTCC
	dCas9_recomb_R	AAGCAGCGTATCCACATAGC
Bisulfite PCR and pyrosequencing	Il33_CpGs_1-4_F	(5’ biotin)TTTAATTTATAAGATTGAAAGTAG AAAATA
	Il33_CpGs_1-4_R	ACTCTAAACCTTTAAAAAAACACTC
	Il33_CpGs_5-6_F	(5’ biotin)TTTGTAATAAGATTTGATATTTTTT TT
	Il33_CpGs_5-6_R	TATTTTATTTTATTCTTTTATTTCTTTCTT
	Il33_CpGs_7-11_F	(5’ biotin)TATTTGTTTTAAAAGTTATATTTAA AAGTT

	Il33_CpGs_7-11_R	ACTATACTTTCCTACAATAAACCCC
	p16_bisPCR_fwd_F	TTTTGATTTTAATTTTTTTGTAAATTT
	p16_bisPCR_fwd_R	TCCCCTTACCTAAAAAAATACC
	p16_bisPCR_rev_F	GGAGGGGTTGGTTGGTTATTA
	p16_bisPCR_rev_R	CTTCTAAAAACTCCCCAAAAAAC
	OFF_TARGET_1_F	(5' biotin)GAAGTTGTTGTTAGTTTAGGAGGT
	OFF_TARGET_1_R	CCCCCTTACAAATAAATTCC
	OFF_TARGET_2_F	(5' biotin)TGTGGTTGAGTAAGTGGTAGATATGTT
	OFF_TARGET_2_R	AATCATCTAATTACCCAAATACACC
	OFF_TARGET_3_F	(5' biotin)GTTTGTTTTTTTTGTGTGGAGAGTT
	OFF_TARGET_3_R	CTACATCATTTACAACCCTAAAACCA
	OFF_TARGET_4_F	(5' biotin)TATTTTTTTTAATTTTTTATTTTTTTAAAT
	OFF_TARGET_4_R	TATATTAATTCCCCAATAATTCTTC
	SERPINB5_bisPCR_F	(5' biotin)TTGTTAAGAGGTTTGAGTAGGAGAG
	SERPINB5_bisPCR_R	CCCACCTTACTTACCTAAAATCACA
	HNF4A_bisPCR_F	TTTTTAAGTGATTGGTTATTTTTTAA
	HNF4A_bisPCR_R	ATATCCCATAACCTCCCAAAACTA

	HNF4A_upstream_bis_F	TTTGGAGTTATAAAATTTAATTTAGGTTG
	HNF4A_upstream_bis_R	AAATAACCAATCACTTAAAAAACCC
	HNF4A_dnstream_bis_F	TAGTTTTGGGAGGTTATGGGATAT
	HNF4A_dnstream_bis_R	ACCCCACCCTCTATAAAATTTTAAA
	TNF_pyro_F	/5biosg/TAGATTGTTATAGAATTTTGGTGGG
	TNF_pyro_R	TTCTATTCTCCCTCCTAACTAATCC
Sequencing primers (pyrosequencing)	Il33_CpGs_1-3	TCCTACTACAAATACTTCTTAAA
	Il33_CpG_5	TCCTCTATAAAACTATATCACAC
	Il33_CpGs_9-11	ACTATACTTTCCTACAATAAACC
	OFF_TARGET_1	AAAACAAACCAAAATAACCAACC
	OFF_TARGET_2	CACCCAATACAAAACTCACACAA
	OFF_TARGET_3	CATCATTTACAACCCTAAAACCAA
	OFF_TARGET_4	TATTAATTCCCCAATAATTC
	SERPINB5_CpGs_1-6	CCCACTACCAACCCAACTCC
	TNFpseqnew1-2	AAAACACCCAAACATCAAAA
	TNFpseqnew3-5	AATAACCCTACACCTCTATC
	TNFpseqnew6-7	AAAACTCTCATTCAACCC
	TNFpseqnew8	AACTTCTACTAACTAACTATACA
	TNFpseqnew9-11	TCTCCCTCCTAACTAATCCCTT
gRNA mutagenesis	SERPINB5_MUT_F	GGCACTCCTCCGGTGTTTCGTCCTT
	SERPINB5_MUT_R	GCCGAGGCGGTTTTAGAGCTAGAA
	gRNATnf2_MUT_F	CCGAGACAGGTTTTAGAGCTAGAAATAGCAAG

	gRNATnf2_MUT_F	TTTCTTCTCCCGGTGTTTCGTCCTTTCC
RT-qPCR (where relevant, h = human, m = murine. h/m = human and murine)	mIl33_001_F	AGAAATCACGGCAGAATC
	mIl33_001_R	GTTGGGATCTTCTTATTTTG
	mIl33_002_F	GCTATTTCCTGTCTGTATTG
	mIl33_002_R	TTCTTTGGTCTTCTGTTG
	h/mGAPDH_F	TGCACCACCAACTGCTTA
	h/mGAPDH_R	GGATGCAGGGATGATGTT
	mACTB_F	GGCTGTATTCCCCTCCATCG
	mACTB_R	CCAGTTGGTAACAATGCCATGT
	allgRNA_F	TGTGGAAAGGACGAAACACC
	allgRNA_R	CGGTGCCACTTTTTCAAGTT
	dCas9_blast_F	GATAAGAACCTGCCCAACGA
	dCas9_blast_R	TTTCTCATTCCCTCGGTCAC
	hmaspin_F	ATAACTGTGACTCCAGGCCC
	hmaspin_R	AGAAGAGGACATTGCCCAGT
	hHNF4A_F	GGCCATGGTCAGCGTGAA
	hHNF4A_R	TTCTGATGGGGACGTGTCGTA
	SL-606_hFMR1_qPCR_F	CAGGGCTGAAGAGAAGATGG
	SL-607_hFMR1_qPCR_R	ACAGGAGGTGGGAATCTGA
	mTNF_qPCR_F	GTAGCCCACGTCGTAGCAAA
	mTNF_qPCR_R	TTGAGATCCATGCCGTTGGC
qPCR	Il33_qChIP_9,10,11_F	CCAAAGTTGTTTAATCTGAGCTACC
	Il33_qChIP_9,10,11_R	GGAAATAGCTGGTCTTGAATGC

	pCpGl_luc_qPCR_F	ACCATTGCCTTCACTGATGC
	pCpGl_luc_qPCR_F	TCCTGTGGTTGGTGTTCAGT
	Actb_gDNA_F	GCCACTCGAGCCATAAAAGG
	Actb_gDNA_R	CAAAAGGAGGGGAGAGGGG
Sanger PCR and sequencing primers	pBABE 3'	ACCCTAACTGACACACATTCC
	TET_mutcheck_F	CTTCTCTGGGGTCACTGCTT
	TET_mutcheck_R	CATCGCAGCCCTCTTCTTTC
	TETmutcheck_seq1	TCGATGGCCCCAGATTTGAT
	TETmutcheck_seq2	ACACCCAAAGAGCGGTTATC

Supplementary Table 2. Names and sequences of oligonucleotide primers used in this study.

SA_TemplateR	AAAAAATCTCGCCAACAAGTTGACGAGATAAACACGGCATTTTGCCTTGTTTTAGTAGATTCTGTTTCCAGAGTACTAAAAC
SP_TemplateR	AAAAAAGCACCGACTCGGTGCCACTTTTTCAAGTTGATAACGGACTAGCCTTATTTTAACTTGCTATTTCTAGCTCTAAAAC
T7FwdAmp	GGATCCTAATACGACTCACTATAG
T7RevAmp_SA	AAAAAATCTCGCCAACAAGT
T7RevAmp_SP	AAAAAAGCACCGACTCGG
SA1_TemplateF	GGATCCTAATACGACTCACTATAGGCAGCATGGAGCCGGCGGCGGTTTTAGTACTCTGG
SA2_TemplateF	GGATCCTAATACGACTCACTATAGAGCAGCATGGAGCCGGCGGCGTTTTAGTACTCTGG
SA3_TemplateF	GGATCCTAATACGACTCACTATAGGAGCAGCATGGAGCCGGCGGGTTTTAGTACTCTGG
SA4_TemplateF	GGATCCTAATACGACTCACTATAGAGCGGGCGGCGGGGAGCAGCAGTTTTAGTACTCTGG
SP1_TemplateF	GGATCCTAATACGACTCACTATAGCAGCATGGAGCCGGCGGCGGTTTTAGAGCTAGAA
SP2_TemplateF	GGATCCTAATACGACTCACTATAGGCAGCATGGAGCCGGCGGCGTTTTAGAGCTAGAA
SP3_TemplateF	GGATCCTAATACGACTCACTATAGAGCAGCATGGAGCCGGCGGGTTTTAGAGCTAGAA
SP4_TemplateF	GGATCCTAATACGACTCACTATAGCGGGCGGCGGGGAGCAGCAGTTTTAGAGCTAGAA

Supplementary Table 3. Primers for *S. aureus*-based strategy for *in vitro* gRNA transcription.

	# CpGs covered	Average Coverage of those CpGs
dCas9Cre_gRNA3_1	40312247	7.1791
dCas9Cre_gRNA3_2	40161948	6.73638
dCas9Cre_gRNA3_3	40126578	6.48471
dCas9Cre_gRNAscr_1	40000845	5.99238
dCas9Cre_gRNAscr_2	40240401	6.7508
dCas9Cre_gRNAscr_3	40248194	6.75135
dCas9TET_gRNA3_1	40182464	6.47765
dCas9TET_gRNA3_2	40718796	9.21089
dCas9TET_gRNA3_3	40309695	6.91347
dCas9TET_gRNAscr_1	40348639	7.0694
dCas9TET_gRNAscr_2	39666464	5.2415
dCas9TET_gRNAscr_3	40026430	6.01691
Untreated_1	40093455	6.09392
Untreated_2	40096678	6.02609
Untreated_3	40123189	6.50818
Average	**40177068.2**	**6.630182**

Supplementary Table 4. Table of all 15 samples on which WGBS was performed, listing the number of CpG-context cytosines covered in each sample and the average read coverage of those cytosines.

	dCas9:gRNAscr			dCas9:gRNA3		
Replicate	**1**	**2**	**3**	**1**	**2**	**3**
Methylation	100	90.90909	71.42857	0	13.33333	11.11111
Reads	4	11	7	7	15	9

Supplementary Table 5. Summary data for coverage and methylation of CpG 9 in the IL33-002 TSS in dCas9:gRNAscr and dCas9:gRNA3 samples.

Off-Target #	Mismatching Target Sequence	Chromosome	Position	Strand	Mismatches
1	CGACaGTGTTAG CaCaCCACAGG	chr17	28576152	+	3
2	CGACTGTGTgAG CTCTgtACTGG	chr14	103608747	+	3
3	CagCTGaGTTAGC TCTCCcCTGG	chr8	30382123	+	4
4	aGACTGTGTTAGt TCcCCAgTGG	chr1	63816122	-	4

Supplementary Table 6. Sequences and locations of predicted mismatched off-target sites for Il33 gRNA3.

Gene	**Ct (average of n=3)**	**Ct (average of n=3)**	**Normalized to Actb**		**Fold Change**
Condition	5-aza 1 μM	Control	5-aza 1 μM	Control	
Tm9sf4	38.74	37.79	2.78658E-07	2.58E-07	1.1
Tnf	31.34	34.205	4.70645E-05	3.1E-06	15.2
Slc44a1	24.045	22.945	0.007391075	0.007599	1.0
Spata2l	28.915	28.38	0.00025275	0.000176	1.4
Cacna1b	29.735	29.69	0.000143168	7.08E-05	2.0
Fam170a	40.295	40.695	9.48353E-08	3.45E-08	2.8
Nars	20.15	20.16	0.109956134	0.052374	2.1
Atp9a	36.025	39.01	1.82965E-06	1.11E-07	16.5
Mkln1	22.535	22.175	0.021050525	0.012958	1.6
Cenpw	24.515	23.22	0.005336095	0.00628	0.8
Fam20a	27.255	26.325	0.000798732	0.00073	1.1
Msi2	23.55	22.595	0.010416396	0.009685	1.1
Taf3	24.16	23.74	0.006824788	0.00438	1.6
Neb	39.28	37.95	1.91653E-07	2.31E-07	0.8
Foxj1	35.96	35.91	1.91397E-06	9.5E-07	2.0
Brinp3	38.435	38.11	3.4426E-07	2.07E-07	1.7
Mcf2	32.795	32.48	1.7167E-05	1.02E-05	1.7
Il33	35.65	35.14	2.37276E-06	1.62E-06	1.5
Actb	16.965	15.905			

Supplementary Table 7. Candidate genes in NIH-3T3 cells for robust induction by 5-aza-2'-deoxycytidine. Expression normalized to *ACTB* and expressed as fold change from water-treated controls. Data is presented as average of three biological replicates (n = 3) each with two technical replicates.

13.1 Oxidized 5mC derivatives and the active DNA demethylation pathway

Chapter 2 of this thesis established dCas9-based steric hindrance as a specific and efficient method for targeted DNA demethylation. This novel technique is relatively simple to use and thus could be adopted by researchers in the DNA methylation field for the assessment of the causal role of specific instances of DNA methylation in gene expression regulation across different physiological processes and research contexts. Chapter 2 also presented several points of evidence to suggest that dCas9-TET systems, on the other hand, are not well-suited for this purpose.

In seemingly filling this fundamental gap in the field and addressing a cardinal misunderstanding of the nature of dCas9-TET systems which, as a result, were inappropriately believed to fill this gap first, other questions are brought to light. In these results, dCas9-TET based systems caused the smallest decreases in methylation but the largest increases in gene expression, while catalytic TET mutants still drove large expression differences in the absence of demethylation. This underscores a critical question in the field: in a physiological context, is it the activity of TET or the completion of the presumed active DNA demethylation pathway that is the aspect more critical to development, differentiation, pluripotency, neuronal activity and other processes? If it is the former, then is it the catalytic or the noncatalytic function of TET? While demethylation by TET enzymes has been reported in different studies covering the full spectrum of these examples, many of these studies demonstrated demethylation by bisulfite sequencing which, due to its inability to discriminate unmethylated cytosine from 5fC and 5caC, may have caused an overemphasis of the importance of true demethylation in these processes. What might instead be the role of 5fC and 5caC in

gene expression regulation? Or, might 5mC oxidation (and ensuing demethylation) be a vestigial function of TET enzymes that only cooccurs with more major effects of the enzymes' noncatalytic activity? These and other questions about the physiological mechanisms and overall significance of the active DNA demethylation pathway still loom.

To be able to investigate these processes, improvements to the sensitivities and specificities of oxidized 5mC derivative sequencing techniques are required. Chapter 3 presents APOBEC-sequencing as a novel method for simple and senstive detection of oxidized 5mC derivatives – particularly 5caC – and demonstrates its application *in vivo* in the mouse brain. Aided by this technique, this chapter also describes a series of genetic manipulations and molecular biology approaches which, when combined with transfected *in vitro* TET2-oxidized reporter plasmids to essentially disentangle TET2 activity from the oxidation-demethylation process, yield new insight into the dynamics of the active DNA demethylation pathway and the regulation of oxidized 5mC derivatives.

Chapter 3: Bisulfite-free sequencing of oxidized cytosines (APOBEC-seq) reveals a causal role of TDG in oxidized promoter re-activation and its ubiquitous MBD3-mediated presence in active gene promoters

Sapozhnikov, D.M., Szyf, M. Bisulfite-free sequencing of oxidized cytosines (APOBEC-seq) reveals a causal role of TDG in oxidized promoter re-activation and its ubiquitous MBD3-mediated presence in active gene promoters. *The EMBO Journal* (under review).

14.1 Abstract

5-methylcytosine oxidation is a fundamental process critical to regulating the epigenetic landscape that determines and maintains cellular identity and physiology. Thymine DNA glycosylase (TDG) is required for efficient demethylation of oxidized promoters, leading to their reactivation, but the dynamics of this process remain unclear. Here, we use APOBEC-seq in tandem with numerous genetic manipulations to dissect the active DNA demethylation pathway and find that the though the TDG catalytic domain exhibits strong binding of oxidized cytosines, the full-length protein is required for the rescue of expression of oxidized promoters in TDG mutants. We report a DNA methylation independent transcriptional activation activity of TDG that confounds the causal relationship between active DNA demethylation and reactivation of oxidized promoters, but find that ultimately TDG-mediated demethylation is not dependent on transcriptional activity or CpG density. We discover that TDG exhibits a surprisingly ubiquitous binding to nearly all active transcription start sites, regardless of promoter RNA polymerase subtype. To understand this binding activity, we employ mass spectrometry and identify MBD3 and all components of the MBD3/NuRD complex as interactors of TDG; knockout of MBD3 reduces TDG binding to several active transcription start sites. We show that MBD3 and TDG co-localize in the genome and that MBD3 is a specific binder of highly oxidized promoter DNA in cells. Finally, we profile DNA oxidation levels genome-wide in human cells and *in vivo* in the mouse brain and find that regions lacking TDG and MBD3 binding exhibit the highest oxidation levels and represent highly tissue-specific genes and enhancers. We speculate that, together with TET family of enzymes, the presence of TDG and MBD3 at transcriptionally active regions may safeguard active genes from spurious CpG methylation and oxidation.

14.2 Introduction

The remarkable plasticity of the functional output from any single genome – the regulation of gene expression – is the fundamental property that enables the development of multicellular organisms with functionally distinct cell types and tissues. DNA methylation constitutes one layer of gene expression regulation in part by directly and indirectly restricting access of transcription factors to gene promoters[465] and the dynamic interplay between DNA methylation and DNA demethylation is critical for various cellular processes, including development, reprogramming, differentiation[219], and responses to cellular cues and its dysregulation can lead to a broad spectrum of diseases as diverse as neurological disease and cancer.

Active DNA demethylation – elucidated only a decade ago[207] – is triggered by the TET family of enzymes, which catalyze the oxidation of methylated cytosines (5mC) through several successive stages of oxidation (5-hydroxymethylcytosine/5hmC, 5-formylcytosine/5fC, and 5-carboxylcytosine/5caC) which can then be removed by glycosylation and base excision repair (BER) pathways[219]. Active DNA demethylation is particularly important in the brain, where nondividing cells must actively regulate DNA methylation levels to modulate gene expression for processes such as learning and memory formation in the absence of passive DNA demethylation by cell division[63,466]. Yet, the study of active DNA demethylation and of the oxidized 5mC derivatives themselves has been hindered by their low abundance, transitory nature, and the destructive and inefficient nature of labeling/detection techniques. Therefore, the relative contributions of active DNA demethylation and these marks to transcription factor binding and gene expression changes remains unclear.

The five MBD protein family members (MECP2, MBD1-4) share a common methyl-CpG binding domain (MBD) and were thus originally characterized as specific 5mC binders, with the exception of MBD3, which lacks four conserved amino acids in its MBD[78]. As the MBD has evolved to directly bind 5mC, it is possible that oxidized 5mC derivatives

could also be recognized by the same binding domain. There is limited and contradictory evidence as to the ability of MBD family proteins to recognize the various oxidized 5mC derivatives, with some studies suggesting an ability of MECP2 and MBD3 to bind oxidized cytosines[263,467], but other studies dispute these findings[364,468-471]. While TDG is known to be involved in the active DNA demethylation pathway[472], little is known about its recruitment to specific regions in the DNA, whether that recruitment is dependent on other oxidation-sensitive transcription factors such as, potentially, the MBD family of proteins, nor about the dependence of TDG activity on local gene expression levels and CpG density and whether DNA demethylation or, perhaps instead, a transcriptional activation activity of TDG, drives the post-demethylation increase in gene activity.

Here, we report a simple and highly efficient bisulfite-free method for the detection of oxidized cytosines – APOBEC-seq – which is commercially available in an optimized formulation and compatible with standard Illumina sequencing technology. Using this method, we study the DNA demethylation dynamics of oxidized transfected promoter-reporter DNA in cells and find that oxidized reporter plasmids are rapidly demethylated in a TDG-dependent manner which can only be rescued by a full-length TDG and not the catalytic domain, despite the latter being sufficient for highly specific binding of 5caC. We report a transcriptional activation capacity of TDG that confounds the causal relationship of active DNA demethylation and gene re-expression and attempt to dissociate the two activities to resolve this question. We further find that the MBD family of proteins binds oxidized CpGs in a reporter plasmid, though knockout of individual MBD proteins in isolation does not cause major differences in expression or demethylation of the reporter plasmid. Genome-wide mapping of ectopic TDG reveals a surprisingly ubiquitous binding activity to active unmethylated and unoxidized promoters despite its strong preference for 5caC binding, which we determine to be partially explained by an interaction with the MBD3/NuRD complex. Finally, we apply APOBEC-

seq to mouse cortices to generate genome-wide oxidized CpG profiles which demonstrate highly tissue-specific oxidation of genes and enhancers.

14.3 Results

Bisulfite-free base-resolution sequencing of oxidized cytosines by enzymatic deamination

In order to study active DNA demethylation of an oxidized transfected promoter-reporter plasmid, we sought to develop a technique to rapidly and efficiently differentiate *in vitro* oxidized DNA from unmethylated DNA that is simple to use and preserves DNA integrity for high quality sequencing. A recent technique developed by New England Biolabs (NEB) and now commercially available[30] introduced the capability to increase 5mC sequencing quality by a two-step non-destructive enzymatic conversion in which TET2 is first used to oxidize 5mC to derivatives that are resistant to subsequent deamination by APOBEC3A, such that only unmethylated cytosines are deaminated, converted to T during PCR, and, using sequencing, produces a readout similar to standard bisulfite conversion (**Figures 1A-B**). We hypothesized that simply skipping the first oxidation step and applying only APOBEC3A would facilitate the direct detection of *in vitro* oxidized cytosines – which are resistant to deamination – and distinguish these marks from 5C and 5mC, which are deaminated by APOBEC3A (**Figure 1C**). Analysis of modified oligonucleotides by APOBEC3A conversion only and (pyro)sequencing (APOBEC-seq) demonstrated complete resistance of 5caC to deamination, partial resistance of 5fC and 5hmC, and full conversion of 5mC and 5C (**Figure 1D**) in agreement with NEB data, though we observed a larger deamination of 5fC and 5hmC than previously reported. In these experiments, APOBEC-seq outperformed MAB-seq (reduced false positive detection) and CAB-seq (reduced false negative detection) in the detection of 5caC. Further testing on unmethylated, *in vitro* methylated (by M.SssI), and *in vitro* oxidized (by TET2) plasmid DNA demonstrated the ability of APOBEC-seq to categorically and quantitatively distinguish oxidized cytosines, including non-CpG

cytosines (Dcm site), from unoxidized (methylated and unmethylated) cytosines (**Figure 1E-F**), though here we observed a slightly reduced conversion of 5mC compared to C. Note that oxidation efficiency of methylated non-CpGs is reduced relative to the oxidation of methyl CpGs (**Figure 1D-E**). Overall, APOBEC-seq specifically detects oxidized cytosines – with a preference towards the highest oxidized form, 5caC – is non-destructive[30,473], and is commercially available in an optimized kit. This allows the simple and efficient analysis of active DNA demethylation of an *in vitro* oxidized promoter-reporter plasmid.

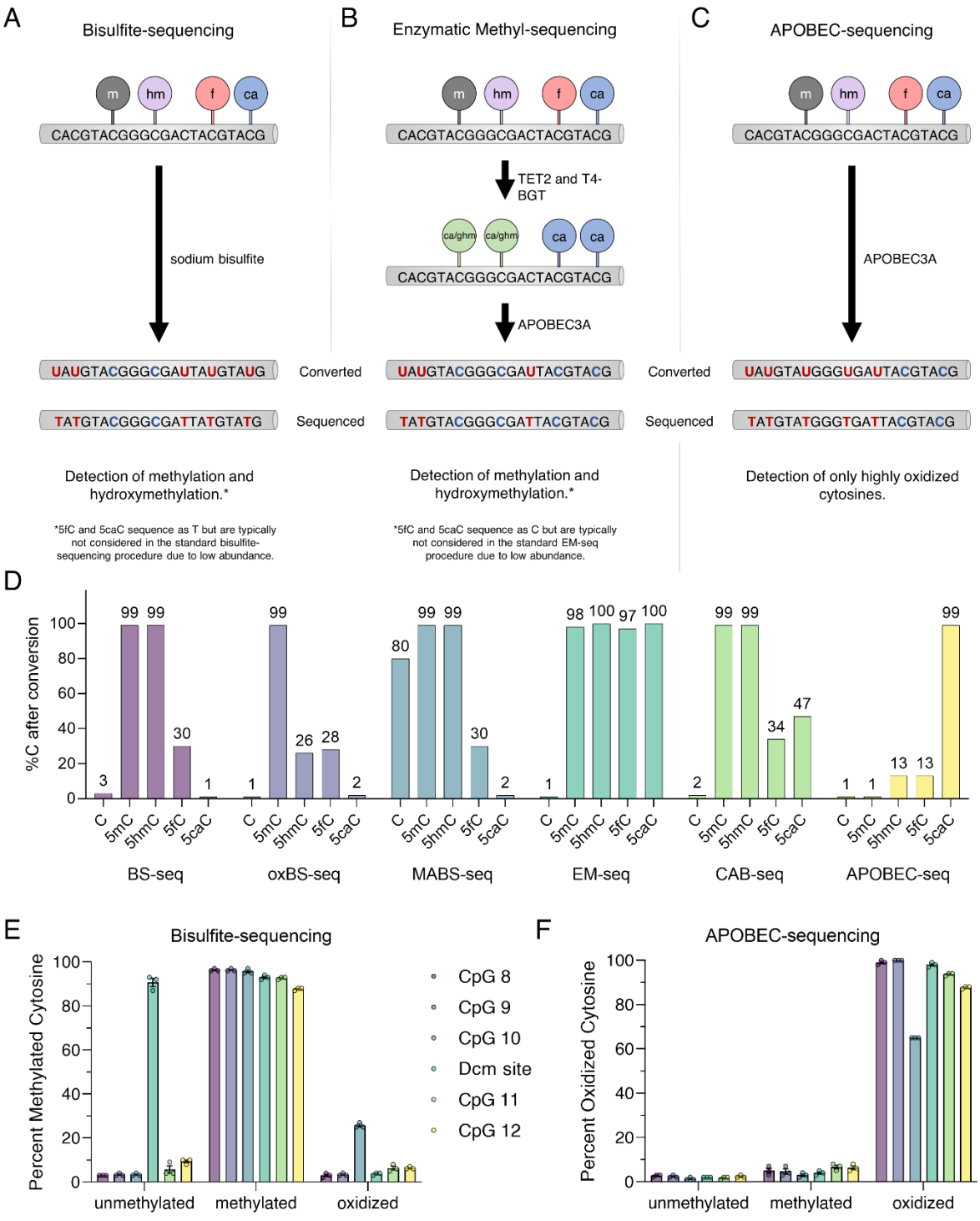

A
Bisulfite-sequencing
m
hm
f
ca
CACGTACGGGCGACTACGTACG
sodium bisulfite
UAUGTACGGGCGAUTAUGTAUG
Converted
TATGTACGGGCGATTATGTATG
Sequenced
Detection of methylation and hydroxymethylation.*
*5fC and 5caC sequence as T but are typically not considered in the standard bisulfite-sequencing procedure due to low abundance.
B
Enzymatic Methyl-sequencing
CACGTACGGGCGACTACGTACG
TET2 and T4-BGT
ca/ghm
ca
APOBEC3A
UAUGTACGGGCGAUTACGTACG
Converted
TATGTACGGGCGATTACGTACG
Sequenced
Detection of methylation and hydroxymethylation.*
*5fC and 5caC sequence as C but are typically not considered in the standard EM-seq procedure due to low abundance.
C
APOBEC-sequencing
CACGTACGGGCGACTACGTACG
APOBEC3A
UAUGTAUGGGUGAUTACGTACG
Converted
TATGTATGGGTGATTACGTACG
Sequenced
Detection of only highly oxidized cytosines.
D
%C after conversion
C
5mC
5hmC
5fC
5caC
BS-seq
oxBS-seq
MABS-seq
EM-seq
CAB-seq
APOBEC-seq
E
Bisulfite-sequencing
Percent Methylated Cytosine
unmethylated
methylated
oxidized
CpG 8
CpG 9
CpG 10
Dcm site
CpG 11
CpG 12
F
APOBEC-sequencing
Percent Oxidized Cytosine
unmethylated
methylated
oxidized

Figure 1. APOBEC conversion and APOBEC-seq for the specific detection of oxidized cytosines. (A-C) Schematic diagram of the conversion steps of two common 5-methylcytosine detection methods – bisulfite-sequencing in (A) and Enzymatic Methyl-sequencing (EM-seq) in (B) – where m = 5-methylcytosine, hm = 5-hydroxymethylcytosine, f = 5-formylcytosine, ca = 5-carboxylcytosine, and ghmC = glucosylated 5-hydroxymethylctosine. (C) APOBEC-seq is presented as a modification to Enzymatic Methyl-sequencing to facilitate the specific detection of highly oxidized cytosines. (D) Average percent cytosine readout by pyrosequencing of three technical replicates of each sequencing technique as a function of CpG modification in a PCR amplicon containing synthesized modified CpG in the primer. The strand containing the modified CpG was specifically amplified after conversion. BS-seq = bisulfite sequencing, OXBS-seq = oxidative bisulfite sequencing, MABS-seq = methylase assisted bisulfite sequencing, EM-seq = enzymatic methyl-seq, CAB-seq = chemical modification-assisted bisulfite sequencing. C = cytosine (unmodified), 5mC = 5-methylcytosine, 5hmC = 5-hydroxymethylcytosine, 5fC = 5-formylcytosine, 5caC = 5-carboxylcytosine. (E) Standard bisulfite sequencing results displayed as percent methylation (C count divided by C + T count) of 5 CpGs and one non-CpG cytosine in the CMV-pCpGl plasmid that is unmethylated, or has been methylated, or methylated and oxidized *in vitro*. (F) Results of the APOBEC-seq workflow on the same plasmids in (E). Data are presented as mean ± SEM.

TDG is required for the reactivation of transfected oxidized promoters in human cells

To study the effects of oxidation on gene expression, unmethylated, methylated, or oxidized reporter constructs (from **Figure 1D-E**) – which encode the CMV promoter driving expression of a luciferase reporter (CMV-pCpGl) – were transiently transfected into 11 different human-derived cell lines. The sequence and CpG structure of the CMV promoter – as well as the SV40 promoter, used below, are presented in **Supplementary Figure 1**. Across all cell lines, methylation of the CMV promoter dramatically decreased luciferase activity whereas the oxidation of the methylated promoter restored high expression, though to a level below that of the unmethylated plasmid (**Figure 2A**). Similar results were obtained with an independent method that used a CMV promoter driving CpGless eGFP expression and measured expression by flow cytometry (**Supplementary Figure 2**).

APOBEC-seq in HEK293 cells revealed that the expression of luciferase activity from the oxidized CMV promoter is accompanied by the replacement of six oxidized cytosines near the transcription start sites (TSS) by unmodified cytosines (hereafter, referred to as “demethylation”) (**Figure 2B**), including a reduced but clear demethylation of a cytosine in a non-CpG context (E. coli Dcm methyltransferase site, C<u>C</u>TGG, underline indicates modified cytosine). Interestingly, expression levels parallel demethylation levels at each time point: for example, oxidized CMV-pCpGl showed 34% demethylation at 12 h and luciferase activity was 35% of unmethylated levels, which became 53% demethylation and 54% expression at 24 h and 71% demethylation and 75% expression at 48h (**Figure 2C**). DNA demethylation of CMV-pCpGl must occur by active (enzymatic) and not passive (dilution) means as there is no detectable replication of the CMV-pCpGl plasmid in HEK293 cells as evidenced by no time-dependent decrease in Dcm site methylation, which cannot be maintained by mammalian

methylation machinery and would thus decrease if any replication had occurred (**Figure 2D**).

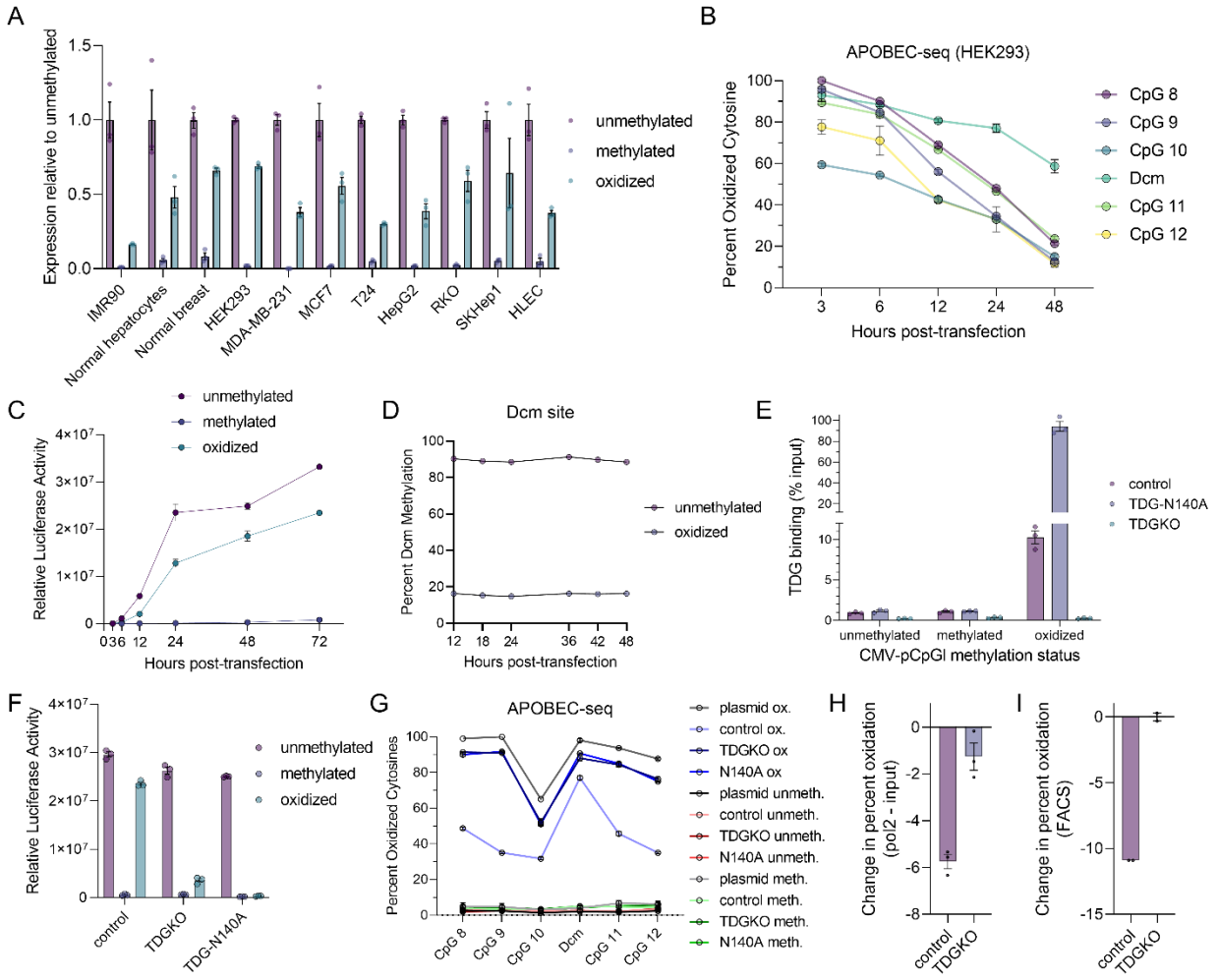


Figure 2. Oxidized DNA is expressed and demethylated in a TDG-dependent manner. (A) Relative luciferase activity of 11 indicated cell lines transfected with 50 ng unmethylated, methylated, or oxidized CMV-pCpGI, normalized to the unmethylated condition in each cell line. (B) Percent oxidation of 5 CpGs and one non-CpG cytosine (Dcm site) in the CMV-pCpGI promoter measured by APOBEC-pyrosequencing at each indicated time point after transfection into HEK293 cells. (C) Relative luciferase activity of 50 ng unmethylated, methylated, or oxidized CMV-pCpGI plasmids transfected into HEK293 cells and collected at indicated time points after transfection. (D) DNA methylation levels of non-CpG Dcm site from bisulfite-pyrosequencing of unmethylated

and oxidized CMV-pCpGl transfected into HEK293 cells, collected at indicated time points post-transfection. (E) Endogenous TDG binding measured by ChIP-qPCR to transfected unmethylated, methylated, or oxidized CMV-CpGl into control HEK293 cells, TDG knockout HEK293 cells (TDGKO), or HEK293 cells with a N140A mutation in endogenous TDG, 48 h post-transfection. Data is normalized to input levels in each sample. (F) Relative luciferase activity as in (C), of 50 ng of three CMV-pCpGl plasmid types 48 h post-transfection into three HEK293 cell lines from (E). (G) Percent oxidized cytosines determined by APOBEC-pyrosequencing of each condition in (F) across 5 CpGs and one non-CpG cytosine (Dcm). Grey line (oxidized plasmid) serves as a reference for original oxidation levels in the plasmid that was used for transfection in all oxidized conditions. Unmethylated and methylated CMV-pCpGl were not assessed by APOBEC-seq; bisulfite-seq data is instead presented in Supplementary Figure 3. (H) Relative enrichment of demethylated CMV-CpGl. Oxidized CMV-pCpGl was transfected into control or TDGKO HEK293 cells and collected 48 h post-transfection by ChIP using antibody against RNA polymerase II. Change in percent oxidation is calculated as the average difference in oxidation percent of the 5 CpGs described in (B) and (G) in pol2-bound DNA and its respective input (total DNA) sample, measured by APOBEC-pyrosequencing. (I) Similar to (H), relative enrichment of demethylated CMV-eGFP-pCpGl, calculated as the average difference in oxidation percent of the 5 CpGs in cells transfected with oxidized plasmid. Here, cells with expression (eGFP+) were collected by fluorescence-activated cell sorting (FACS) and CMV promoter oxidation was quantified by APOBEC-pyrosequencing. Difference is expressed as oxidation percent in unsorted cells subtracted from oxidation percent in eGFP+ cells. For all figures, data are presented as mean ± SEM, with individual replicates plotted as circles in all bar graphs; in line plots, circles represent the mean and n=3 biological replicates across all experiments except for n=2 in (I). Control cells used for all experiments were clonal cell lines prepared and processed as TDGKO cells, but were originally transfected with Cas9 and a scrambled gRNA.

As determined by chromatin immunoprecipitation followed by quantitative PCR (ChIP-qPCR), thymine DNA glycosylase (TDG) binds at high levels exclusively to oxidized CMV promoter (**Figure 2E**); this binding can be eliminated by TDG knockout (TDGKO) by CRISPR/Cas9 (**Figure 2E**, **Supplementary Figure 3A-D**) or, conversely, can be augmented by introducing a mutation in endogenous TDG (TDG-N140A, **Supplementary Figure 3A,E**) which preserves DNA binding ability but abolishes its catalytic activity[474]. TDGKO cells show a dramatic reduction in expression of oxidized CMV-pCpGl (**Figure 2F, Supplementary Figure 2**) – though it remains expressed at a level slightly higher than methylated CMV-pCpGl – and mutagenesis of the catalytic amino acid (N140A) of TDG suppresses expression further than in TDGKO, likely due to the accumulation of mutant TDG on the DNA and interference with transcriptional processes.

Paralleling luciferase assay results, ChIP-qPCR demonstrated that RNA polymerase II (pol2) binds unmethylated CMV promoter to a greater degree than methylated promoter regardless of TDG mutational status; pol2 binding to the oxidized promoter remained high in control cells but was reduced to levels comparable to the methylated condition in TDGKO and TDG-N140A cell lines (**Supplementary Figure 4A)**. Similar patterns of binding were observed for pol2 phosphorylated on serine 5 (pol2-PS5), the form of pol2 associated with active initiation of transcription (**Supplementary Figure 4B**). Reduction of luciferase expression in TDGKO and TDG-N140A conditions was concurrent with absence of demethylation (**Figure 2G**) at 24 hours post-transfection though, consistent with the slightly increased expression of oxidized compared to methylated CMV promoter in TDGKO, a small demethylation was still observed. There were no changes in DNA methylation of unmethylated or methylated plasmid in any condition (**Supplementary Figure 4C**). APOBEC-seq of DNA bound by pol2 revealed an enrichment for unmethylated cytosines compared to that of input, suggesting that transcription preferentially occurs from oxidized promoters that are demethylated (**Figure 2H**). We further probed the demethylation of all 35 CpGs and 3 non-CpGs in

the CMV promoter and found the entire oxidized promoter to be demethylated (**Supplementary Figure 4D**) with comparable demethylation of plus and minus strands (**Supplementary Figure 4E**). To test whether replacement of oxidized cytosines is dependent on the extent of promoter activity also tested the weaker SV40 promoter: the expression from this CpG-poor promoter is less inhibited by DNA methylation (**Supplementary Figure 4F**) but nevertheless, the oxidized cytosines are replaced by unmethylated cytosines as they are in the stronger CMV promoter (**Supplementary Figure 4G)**. Replacement of oxidized cytosines in SV40-pCpGl and CMV-pCpGl by unmethylated cytosines was replicated using MABS instead of APOBEC-seq (**Supplementary Figure 4H-I**).

Finally, we tested whether oxidized CpGs could be maintained through cell division and DNA replication. Plasmids were transfected and integrated into the genome (stable transfection) by co-transfection with 10X (molar) less plasmid expressing the puromycin resistance gene followed by selection with puromycin. Stable plasmid integration was verified after 20 days of selection as evidenced by detection of pCpGl by qPCR (**Supplementary Figure 4J**) and complete demethylation of the Dcm site (**Supplementary Figure 4K**). While a portion of the methylation of stably integrated methylated SV40-pCpGL (**Supplementary Figure 4L**) and CMV-pCpGl (**Supplementary Figure 4M**) could be maintained after 20 or 40 days of selection there was no observable oxidation maintained at 20 days of selection for stably transfected oxidized SV40-pCpGl (**Supplementary Figure 4N**) or CMV-pCpGL (**Supplementary Figure 4O**), even in TDGKO cells, suggesting there are no mechanisms – at least in HEK293 cells – to maintain oxidized states of cytosines and that oxidation is lost through passive mechanisms irrespective of the presence of TDG. Furthermore, the methylated state was maintained more efficiently in TDGKO cells than in control cells (by 22-33%), suggesting that TDG plays a role in the demethylation of stably integrated methylated plasmid DNA (**Supplementary Figure 4M**).

Full length TDG but not its catalytic domain can restore demethylation and gene expression

Next, we studied the effects of rescue of TDGKO cells with exogenous TDG. We found that the mutant catalytic domain of TDG (TDG-N140A-CD) exhibits a remarkable binding to oxidized CMV promoter that is orders of magnitude higher than wild type TDG and several fold higher than the mutant full-length TDG (TDG-N140A-FL) (**Figure 3A**). In fact, this strong and selective binding could be used to purify (by ChIP) completely oxidized DNA from transfected mixtures containing 1:1 or 99:1 ratios of unmethylated:oxidized DNA (**Figure 3B**), which may have future applications in enrichment and capturing of oxidized DNA. Wild-type TDG-CD and TDG-FL still bound oxidized DNA but at a marked reduced capacity compared to counterparts with the N140A mutation (**Figure 3A**), likely because of dissociation of the wild-type TDG after completing its enzymatic activity.

Interestingly, despite the strong binding of TDG-N140A-CD to oxidized DNA, oxidized promoter expression (**Figure 3C**) and demethylation (**Figure 3D**) in TDGKO cells could only be restored by rescue with wild-type TDG-FL but not with wild-type TDG-CD. In contrast, rescue with either of the mutant TDG constructs (TDG-N140A-CD and TDG-N140A-FL) suppressed oxidized CMV promoter expression and demethylation to levels below that of control TDGKO cells, consistent with our previous results in **Figure 2D**. Similar results were obtained in control cells, where TDG-N140A-CD, TDG-N140A-FL, and TDG-CD markedly repressed oxidized CMV-pCpGl expression and even TDG-FL interfered with expression (**Figure 3E**), possibly by suboptimal activity of the transgenic construct and interference with the more efficient endogenous TDG. These data demonstrate that the catalytic domain of TDG is sufficient for binding to oxidized transfected DNA but the full protein is required to initiate replacement of the oxidized base with cytosine and for the resulting reactivation of expression. A possible explanation is that SUMOylation at the C-terminus of TDG (which is absent in CD

constructs) is required to facilitate release of TDG from the DNA after enzymatic activity[475,476]. Alternatively, the lack of activity of wild-type TDG-CD may be caused by its >10-fold preference for glycosylation of 5fC compared to 5caC, which is the main modification in our reporter plasmid[474].

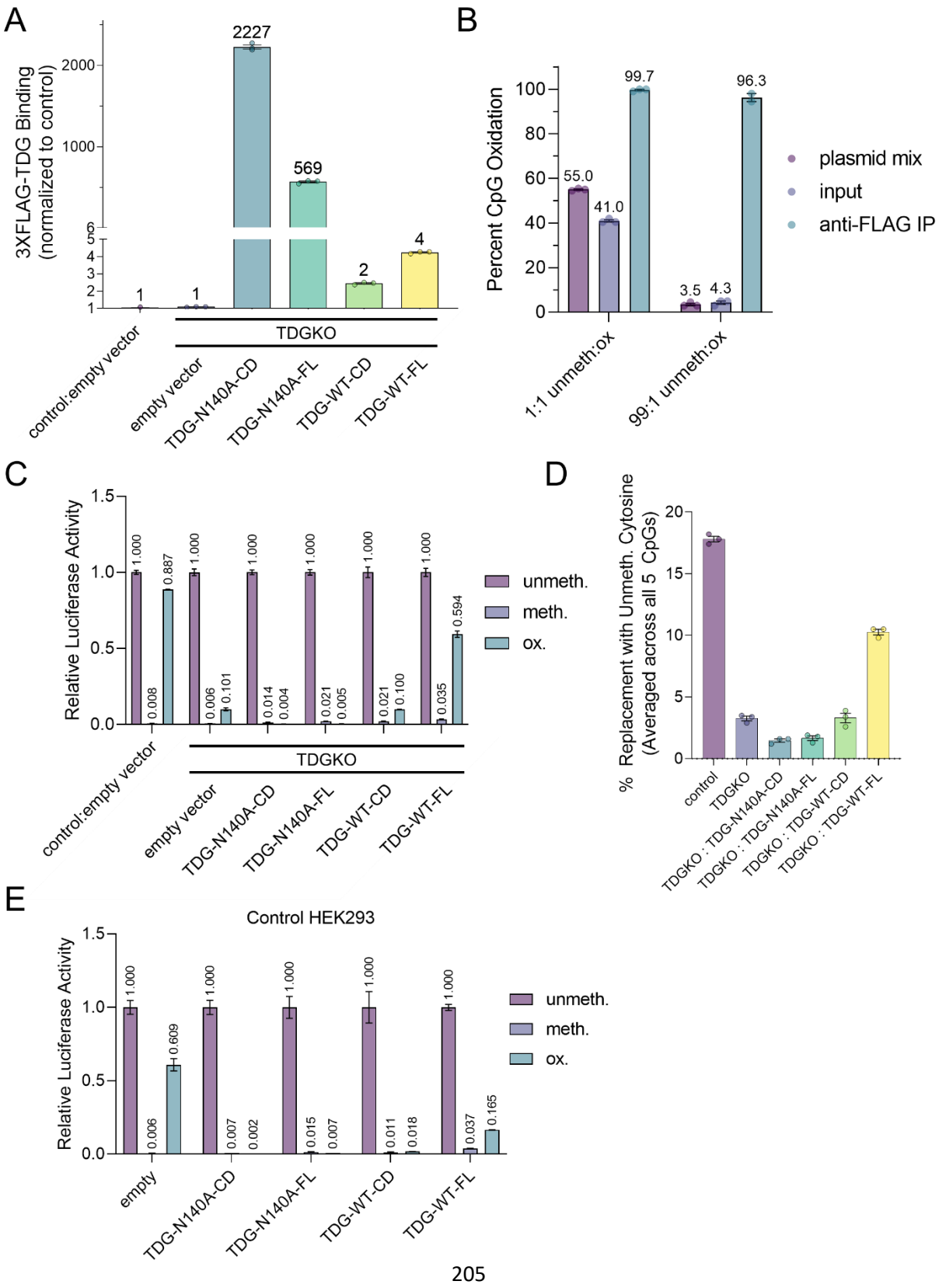

A
3XFLAG-TDG Binding (normalized to control)
2227
569
4
2
1
1
TDGKO
control:empty vector
empty vector
TDG-N140A-CD
TDG-N140A-FL
TDG-WT-CD
TDG-WT-FL
B
Percent CpG Oxidation
99.7
96.3
55.0
41.0
3.5
4.3
plasmid mix
input
anti-FLAG IP
1:1 unmeth:ox
99:1 unmeth:ox
C
Relative Luciferase Activity
unmeth.
meth.
ox.
1.000
0.008
0.887
0.006
0.101
0.014
0.004
0.021
0.005
0.021
0.100
0.035
0.594
TDGKO
control:empty vector
empty vector
TDG-N140A-CD
TDG-N140A-FL
TDG-WT-CD
TDG-WT-FL
D
% Replacement with Unmeth. Cytosine (Averaged across all 5 CpGs)
control
TDGKO
TDGKO : TDG-N140A-CD
TDGKO : TDG-N140A-FL
TDGKO : TDG-WT-CD
TDGKO : TDG-WT-FL
E
Control HEK293
Relative Luciferase Activity
unmeth.
meth.
ox.
1.000
0.006
0.609
0.007
0.002
0.015
0.007
0.011
0.018
0.037
0.165
empty
TDG-N140A-CD
TDG-N140A-FL
TDG-WT-CD
TDG-WT-FL

Figure 3. Effects of TDG rescue. (A) Binding to oxidized CMV-pCpGl of various forms of 3X-FLAG-tagged TDG stably expressed in HEK293 cells. Empty vector indicates control cell lines that express the empty (no TDG) lentiviral construct that all TDG forms were cloned into. Binding was calculated by ChIP-qPCR and normalized to input and is shown as relative to control cells with empty vector. (B) Percent oxidation of CpG 8 in CMV-pCpGl in mixtures of unmethylated and oxidized CMV-pCpGl plasmids at indicated ratios, in pure plasmid mix or after transfection (48 h) into HEK293 cells expressing 3XFLAG-TDG-N140A-CD and ChIP using an anti-FLAG antibody or input (total DNA) control from the two conditions. CpG 8 is depicted as it is the first CpG sequenced in the pyrosequencing assay and is the least technically variable; pyrosequencing quality decreases with sequencing length. Total DNA transfected (500 ng) was the same across both conditions. (C) Effects of stable rescue with various forms of TDG in TDGKO cells as compared to control HEK293 cells on relative luciferase activity of three forms of CMV-pCpGl 48 h post-transfection, normalized to the unmethylated condition in each cell line to avoid any variability stemming from error in cell counting across cell lines when plated for transfection. (D) Percent of active DNA demethylation calculated by APOBEC-pyrosequencing in conditions in (C). (E) Same as (C) but on a control (not TDGKO) genetic background.

Relationship between transcription and TDG-mediated demethylation

Despite the accepted paradigm that TDG-mediated replacement of oxidized cytosines with unmethylated cytosines leads to gene reactivation, it remains unclear whether loss of oxidation leads to transcription or whether transcription leads to DNA demethylation. In fact, all forms of overexpressed ectopic TDG, but particularly full length TDG (regardless of mutational status), exert a trans-activation effect even on unmethylated (**Supplementary Figure 5A**) or methylated DNA (**Supplementary Figure 5B**). Therefore, transactivation by TDG can occur independently of any changes to DNA methylation status. It was previously shown that the DNA demethylation triggered by induction of histone acetylation with trichostatin A (TSA) was dependent and preceded by transcriptional activation[477]. To test whether this is the case in our model, we assessed the extent of demethylation of the oxidized promoter in the presence of pharmacological inhibitors of transcription. While the transcriptional inhibitors actinomycin D and DRB reduced demethylation (**Supplementary Figure 5C**) α-amanitin failed to do so. The inhibition of demethylation in presence with actinomycin D or DRB did not appear to be an indirect effect of reduced TDG transcription as there was no decrease in TDG protein levels (**Supplementary Figure 5D**). Similarly, cycloheximide – a translational inhibitor – also reduces demethylation but likely through reduction of TDG protein (**Supplementary Figure 5D**) and accordingly reduced recruitment of TDG to DNA (**Supplementary Figure 5E**).

Still, interpretation of the results obtained with global transcriptional inhibitors may be confounded by difficulties in protein level normalization that do not allow confident determination that there is no reduction in TDG protein. In addition, altered cell survival induced by these inhibitors (**Supplementary Figure 5F,G**) might have indirect effects. The fact that α-amanitin does not inhibit demethylation further reduces the confidence in concluding that transcription is required for demethylation.

Therefore, to assess the role of transcription in TDG-triggered demethylation by an independent method, we reduced the transcription capacity of the oxidized CMV-pCpGl plasmid in *cis* by mutating the TATA box of the CMV promoter (ΔTATA). This mutation reduced transcription 38-fold (**Supplementary Figure 5H**) but had no effect on demethylation extent or rate (**Supplementary Figure 5I**). However, this promoter still drives considerable gene expression, which may still explain the observed demethylation. We therefore generated reporter plasmids containing only the most 3' CpG of the CMV promoter. There was no detectable gene expression from this reporter (**Supplementary Figure 5J**). Although the single oxidized CpG in this promoter was demethylated (**Supplementary Figure 5K**), demethylation was reduced relative to CMV-pCpGl. To determine whether this reduction in demethylation was a consequence of inhibition of transcription or consequence of low CpG content, we introduced a CpG-free CMV promoter (all CG dinucleotides mutated to TG) upstream of the single CpG. While this promoter was expressed, demethylation was still reduced to a similar extent, suggesting that reduced demethylation of the single CpG promoter is not a consequence of reduced transcription but most probably a result of lower CpG-density. Nevertheless, the fact that a non-expressed single CpG promoter was demethylated confirms that demethylation by TDG is not mediated by increased transcription and that high CpG density is not a prerequisite for demethylation by TDG but that these factors might be contributing to increase the efficacy of demethylation.

Does TDG induce expression through demethylation? To study this question, it is necessary to dissociate TDG-mediated activation of transcription and demethylation. We attempted to do this by disrupting a downstream step in TDG-mediated DNA demethylation. The current understanding is that the demethylation pathway is triggered by the glycosylase activity of TDG on an oxidized CpG substrate, which results in the generation of an apyrimidinic (AP) site that then must be cleaved by AP endonuclease 1 (APEX1). APEX1 is thought to be required for the base excision repair pathway and knockout of APEX1 has been shown to lead to complete loss of AP site cleavage

activity in HEK293 cells[478]. Therefore, APEX1 knockout should prevent demethylation while leaving the TDG transactivation activity intact. However, surprisingly, complete knockout of APEX1 by CRISPR/Cas9 (**Supplementary Figure 5L,M**) only produced a mild reduction in oxidized CMV-pCpGl demethylation (**Supplementary Figure 5N**) and expression (**Supplementary Figure 5O**). Additionally, APEX1 knockout reduced demethylation further on a TDG knockout background, suggesting APEX1 can mediate a more minor DNA demethylation activity by a non-TDG pathway, but is not required for TDG-mediated DNA demethylation. Therefore, this approach could not dissociate demethylation from transcriptional activation, and we were unable to fully address the question of whether activation of transcription of oxidized promoters is triggered by TDG-mediated demethylation. A possible explanation for the surprising finding that demethylation by TDG occurs even in the absence of APEX1 is that there is an alternative pathway for base excision repair following TDG glycosylase activity. There is evidence that NEIL proteins can directly substitute for APEX1 activity in TDG-mediated active DNA demethylation[216]. Complete knockout of these proteins in addition to APEX1 might be required to confidently dissociate transcriptional activation from demethylation.

The MBD proteins bind oxidized DNA; MBD3 is a specific binder

Our model of transfected unmethylated, methylated, or oxidized forms of CMV-pCpGl combined with TDG knockout – which stabilizes the oxidized state – allows us to uniquely profile in a quantitative manner the interactions of each MBD family member with each epigenetic state in live cells. ChIP-qPCR using antibodies specific to each MBD protein confirmed significant binding preference to methylated over unmethylated DNA for MBD1, MBD2, MBD4, and MECP2 (**Figure 4A**) and not for MBD3, consistent with the literature. Control ChIP experiments examining TATA-binding protein (TBP) demonstrated the expected inverse binding pattern characterized by inhibition of binding by DNA methylation[128], as well as by oxidation (**Figure 4A**).

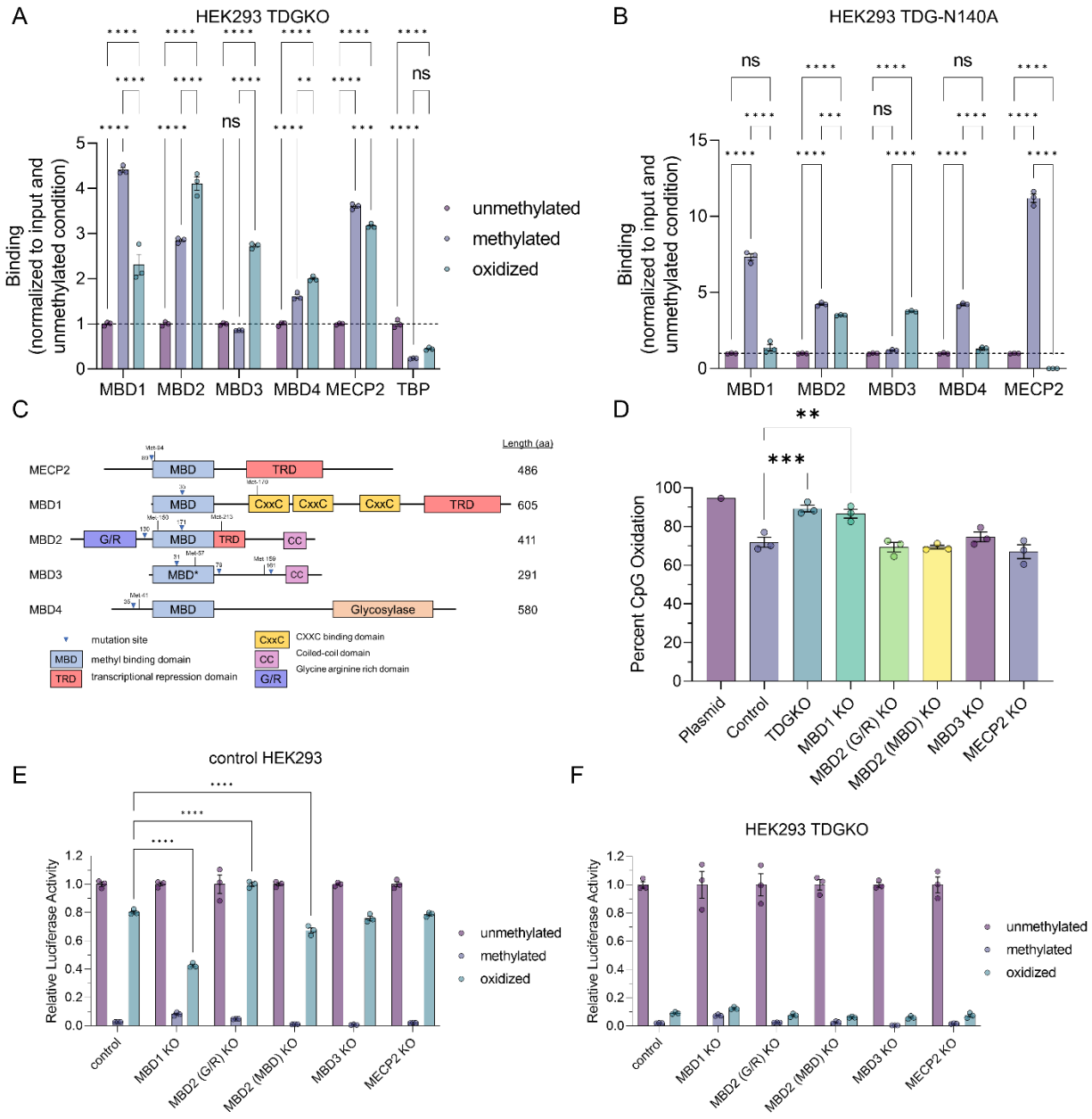


Figure 4. The MBD family of proteins interact with oxidized DNA. (A) Binding of each MBD family member and TBP to unmethylated, methylated, or oxidized CMV-pCpGl as determined by ChIP-qPCR. Data is normalized to input for each sample and to the unmethylated condition within each protein. 500 ng each plasmid was transfected into HEK293 TDGKO cells. (B) Same as (A), but performed in HEK-TDG-N140A cells.

(C) A schematic diagram of the domain structure of MBD family proteins. CRISPR/Cas9 gRNA target sites are indicated with blue triangles and labeled with the amino acid number within the protein sequence. For each gRNA, the most immediate 3' alternative start codon is marked "Met" and with its amino acid number to demonstrate the potential expression of unmutated alternative isoforms despite successful frame-shift mutations by CRISPR/Cas9. For MBD2, the gRNA targeting amino acid 130 was used to generate a G/R domain-less MBD2, while translation from Met-150 produced an MBD2 isoform which retained the MBD domain and all C-terminal sequences, which could still be detected by western blot. A second mutation by CRISPR/Cas9 was introduced on the G/R domain-less background at amino acid number 171 to generate cells with a disrupted MBD (of MBD2) domain. For MBD3, CRISPR/Cas9 mutagenesis with the gRNA targeting amino acid 31 still yielded a largely intact protein detectable by western blot, and gRNAs targeting amino acids 79 and 161 were co-transfected on this background to produce cells with complete MBD3 knockout. For MBD4, the CRISPR/Cas9 mutagenesis strategy failed due to considerable translation from 3' Met-41 and MBD4 knockout was not further pursued. (D) Percent CpG oxidation determined by APOBEC-pyrosequencing of oxidized CMV-pCpGl plasmid transfected into control cells, TDGKO cells, or cells with various MBD knockouts. The "plasmid" condition indicates oxidation levels of the original untransfected plasmid. CpG oxidation was averaged across 4 CpGs in the CMV-pCpGl promoter pyrosequencing assay. (E) Relative luciferase activity, normalized to total protein content and to the unmethylated condition in each cell line, of unmethylated, methylated, or oxidized CMV-pCpGl transfected into each indicated knockout cell line on a control HEK293 background. (F) Same as (E) except performed on a HEK293 TDGKO knockout background, where "control" indicates HEK293 TDGKO cells. For (A-B), p-values were calculated with the Tukey's multiple comparisons test to compare all plasmids types within each cell line. For (D-F), p-values were calculated with Dunnett's multiple comparisons test using control HEK293 cells (D) or oxidized CMV-pCpGl expression in control cells (E,F) as the control condition. * indicates $p<0.05$, ** indicates $p<0.01$, *** indicates $p<0.001$, ****

indicates $p<0.0001$, after multiple correction, and ns indicates no statistically significant difference. In cases where no comparison and p-value are plotted in (D-F), the differences were not statistically significant. All tests were performed using GraphPad Prism v9.4.1.

We were surprised to discover that all MBD proteins exhibited significantly stronger binding to oxidized CMV-pCpGl than fully unmethylated plasmid. However, only MBD2 and MBD3 bound oxidized DNA at higher levels than methylated DNA: here, it is important to note that the MBD2 antibody may be cross-reactive to MBD3 (MBD3 antibody is specific to MBD3 only). The fact that MBD3 binds oxidized DNA but does not distinguish methylated from unmethylated DNA classifies it as a specific binder of oxidized DNA. To confirm the specificity of this binding assay we introduced competitive binding by TDG-N140A: we repeated the ChIP experiments in HEK293 cells with the TDG-N140A mutation, where we have shown that TDG-N140A strongly binds oxidized DNA and its inability to trigger glycosylation prevents its dissociation from the DNA to a degree that outcompetes binding of other proteins (endogenous TDG). In this system, all MBD family proteins retained their specificity for methylated DNA (**Figure 4B**). However, binding of MECP2 to oxidized DNA was reduced to completely undetectable levels and binding of MBD1 and MBD4 to oxidized DNA were reduced to that of unmethylated levels, which may represent background levels. Interestingly, MBD2 and MBD3 were still able to bind oxidized DNA despite obstruction by TDG-N140A: this suggest that MBD2 and MBD3 do not compete with TDG for binding to oxidized DNA and is consistent with the possibility that MBD3 or MBD2 recruit TDG to oxidized DNA, which we tested below.

To assess whether the MBD proteins may therefore influence the expression and demethylation of oxidized DNA, we used CRISPR/Cas9 to individually knockout each MBD family member in HEK293 cells and measured the effect of each knockout on CMV-pCpGl expression and oxidation levels. We produced knockouts MBD family members (except MBD4), and an additional specific deletion of only the MBD2 isoform containing the G/R domain (**Figure 4C, Supplementary Figure 6**). The only significant effect on demethylation of transfected oxidized CMV-pCpGl was detected in MBD1 knockout cells, which resulted in a reduced demethylation of the promoter (**Figure 4D**). Broadly, we found that the effects of knockout of any individual MBD family member on

luciferase activity were minor. There were no significant changes in methylated promoter activity compared to unmethylated promoter activity in control HEK293 cells (**Figure 4E**) or TDGKO cells (**Figure 4F**) as a function of any MBD protein knockout.

Though contrary to the paradigm of suppression of methylated promoter expression by MBD proteins, the lack of de-repression of methylated DNA by knockout of any individual MBD family member is consistent with recent findings[132]. Likewise, there were no changes in oxidized promoter activity in TDGKO cells, but a small significant decrease in expression with MBD1 knockout – which parallels demethylation findings in **Figure 4D** – as well as in the MBD-containing MBD2 isoform knockout cell line, and a small increase in expression in cells lacking the G/R domain of MBD2.

TDG binds nearly all active gene promoters transcribed by RNA polymerase I, II, and III

Next, we performed ChIP-sequencing (ChIP-seq) with anti-FLAG antibody in HEK293 cells stably expressing 3XFLAG-TDG-N140A-CD in an effort to identify the binding profile of TDG and maximize detection of oxidized CpGs by employing the mutant form. We discovered 16,468 significant 3XFLAG-TDG-N140A-CD peaks replicated across all three biological replicates and 21,907 peaks replicated across at least two samples. Parallel experiments with negative control samples which did not express 3XFLAG-tagged TDG-N140A-CD identified 245 artifactual peaks. We report a striking binding profile of TDG-N140A-CD in the promoters of transcriptionally active genes: an example genome browser view of the *TP53* promoter region demonstrates TDG-N140A-CD peaks in the *TP53* promoter but also distinct peaks at two alternative promoters of the neighboring *WRAP53* gene which exhibit RNA polymerase II (pol2) binding (**Figure 5A**). Importantly, TDG is absent from at least one alternative *WRAP53* transcription start site which lacks significant pol2 binding. This behavior was consistent genome-wide – with nearly 99% (15,174 of 15,476) of active promoters (those with significant pol2 peaks)

also containing significant TDG peaks compared to only 2.6% (205 of 7,803) of high-confidence inactive promoters (no pol2 peaks and not detected by RNA-seq). Five active gene promoters were selected for validation by ChIP-qPCR and, like FLAG-tagged TDG-N140A-CD, endogenous TDG – with or without the N140A mutation – exhibited significant binding in these promoters (**Supplementary Figure 7A-E)**.

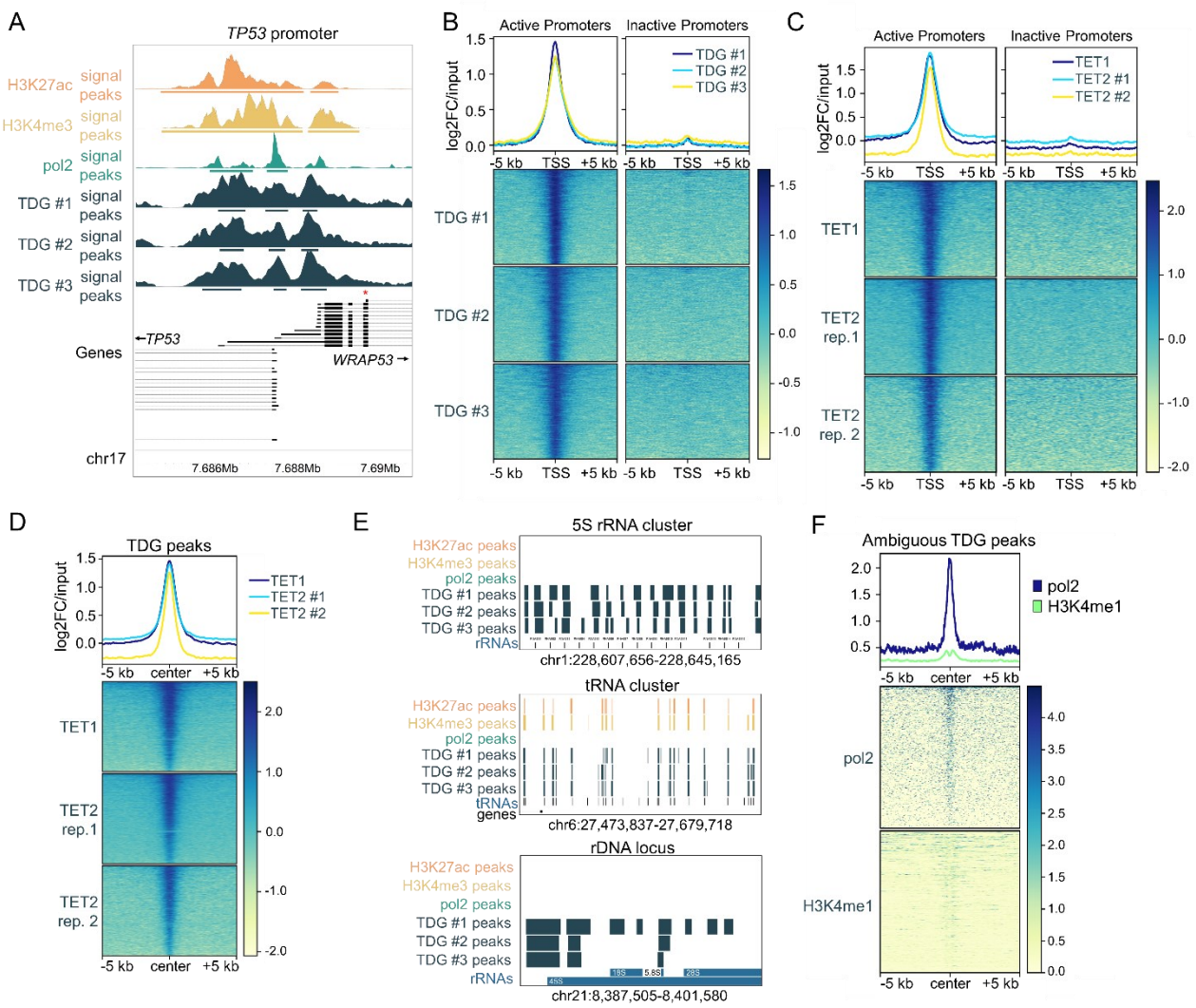


Figure 5. ChIP-seq of 3XFLAG-TDG-N140A-CD. (A) A genome browser view of the *TP53* and *WRAP53* promoter region, with tracks depicting public H3K27ac, H3K4me3, RNA polymerase II (pol2) data signal, and three replicates of 3XFLAG-TDG-N140A-CD signal generated in this study, each plotted alongside statistically significant peaks of binding. The red asterisk marks an alternative *WRAP53* TSS that does not exhibit pol2 binding, active histone marks, or TDG binding. (B) Genome-wide signal data from (A) for three TDG ChIP-seq replicates, plotted across the TSS ± 5 kb of active promoters (those with significant pol2 peaks from public data) and high-confidence inactive promoters. The signal value corresponds to TDG signal normalized to input (log2 scale) for each sample (log2FC/input = log2 fold-change over input). The top panel averages the log2FC/input signal across all active or inactive promoters while the bottom panels show signal over individual promoters. (C) Same as (B) but using public data for a single replicate of TET1 ChIP-seq or two replicates of TET2 ChIP-seq. (D) Same data as (C) but plotted over significant 3XFLAG-TDG-N140A-CD peaks. (E) Significant peaks of all tracks from (A) plotted across a cluster of 5S rRNA genes (top), a cluster of tRNA genes (middle) or a representative rDNA (*RNA45S1*) locus (bottom). (F) RNA polymerase II (blue) and H3K4me1 (green) signal from public data plotted over ambiguous TDG peaks (>5 kb from the nearest gene and not marked by significant H3K4me1/H3K4me3/H3K27ac or pol2 peaks).

TDG signal peaked specifically at the TSS of active genes (**Figure 5B)**. Reciprocally, 8,702 of the 21,907 (40%) of replicated TDG peaks occurred in active promoters compared to 182 (0.8%) at high-confidence inactive promoters. Furthermore, 12,523 (57%) of TDG peaks occurred in dually marked H3K4me3/H3K27ac regions – which represents 78% of all H3K4me3/H3K27ac regions – and an additional 4,044 (18%) and 1,635 (7%) TDG peaks occurred in H3K4me3 or H3K27ac peaks alone, respectively. 3,558 (16%) TDG peaks were in H3K27ac regions that were not in gene promoters, suggesting a further involvement of TDG in gene enhancers: **Supplementary Figure 7F** shows an example of a typical intergenic enhancer which is bound by TDG – it is marked by H3K27ac/H3K4me3 and shows increased pol2 binding and production of small enhancer RNAs. In total, despite only 40% of TDG peaks occurring in promoters, nearly 80% (17,381) of TDG peaks were found in sequences marked by transcriptionally active histone marks (H3K4me1/H3K27ac/H3K4me3). Like TDG-N140A-CD, TET1 and TET2 – which catalyze the first step of active DNA demethylation – also show striking binding profiles to active and not inactive TSS (**Figure 5C**). 14,698 (67%) of TDG-N140A-CD peaks were also sites of significant TET1 or TET2 peaks (**Figure 5D**) (reciprocally, 13,038 of 40,918 TET1 peaks and 11,673 of 24,049 TET2 peaks). Interestingly, the majority of overlapping significant TET1/TET2 peaks (10,013 of 13,069 or ~77%) were also bound by TDG-N140A-CD, possibly implying a ubiquitous presence of active DNA demethylation machinery which safeguards active gene promoters from hypermethylation.

Still, a large fraction of replicated TDG peaks were found in regions that do not contain gene promoters and are not characterized by any specific histone modifications. We further inspected these regions manually. Unexpectedly, TDG peaks were consistently present across 5S rRNA genes – which are transcribed by RNA polymerase III and lack H3K4me3/H3K27ac peaks – (**Figure 5E**) and this was validated by ChIP-qPCR

(**Supplementary Figure 7C**). Like 5S rRNA genes, tRNAs are also transcribed by RNA polymerase III: distinct TDG peaks were present in tRNA genes, such as those within the largest human tRNA cluster on chromosome 6 (**Figure 5E**). These tRNA genes, however, did exhibit H3K4me3/H3K27ac modifications typical of active promoters. More broadly, 222 of 619 human tRNA genes contained TDG peaks: all 222 were expressed according to small-RNA-seq data. Other small non-coding RNA genes transcribed by RNA polymerase III – including *snaR* clusters on chromosome 19 (e.g., *SNAR-A4*), YRNA genes (e.g., *RNY3*), and the BC200 RNA (*BCYRN1*) – also contained significant TDG peaks. RNA polymerase I is responsible for the transcription of rDNA repeats which encode the 45S rRNA, the precursor for 18S, 5.8S, and 28S rRNAs. The rDNA TSS also contains TDG peaks (**Figure 5E**). Together, these data reveal a nearly ubiquitous presence of TDG in active gene promoters across all promoter subtypes. TET1 and TET2 peaks were also widely distributed throughout all examples of non-coding RNA genes (tRNAs, 5S rRNAs, 45S rRNA, etc.). Finally, we defined an additional 1,942 ambiguous replicated TDG peaks that were >5 kb from the nearest gene and were not marked by significant H3K4me1/H3K4me3/H3K27ac or pol2 peaks: however, these peaks exhibited increased pol2 binding than surrounding regions and a bimodal H3K4me1 distribution typical of transcriptionally active or poised regions (**Supplementary File 1, Figure 5F**), suggesting a regulatory role of TDG in these regions. Though 72% of TDG peaks contained CpG islands, this is equivalent to the CpG island rate across all promoters (72%)[96] and this, combined with a large fraction of TDG peaks in CpG-poor regions (**Supplementary Figure 7G**), suggests no preference of TDG for CpG islands. To confirm that the binding of TDG at promoters was not an artifact specific to HEK293 cells, we also profiled by ChIP-seq the binding of TDG-N140A-CD in human liver (HepG2) and breast (MCF-7) cancer cell lines, as well as in a noncancerous and untransformed primary human fetal lung fibroblast cell line (IMR90). In all three cell lines, promoters were the most significantly enriched genetic element, representing 219/7,192 ($p = 2.11 \times 10^{-80}$), 1,569/18,084 ($p = 0$), and 1,017/21,638 ($p = 0$) of significant TDG-N140A-CD peaks in HepG2, MCF-7, and IMR90 cells,

respectively (**Supplementary File 2**). Furthermore, we identified several cell-type specific genes with significant 3XFLAG-TDG-N140A-CD peaks in promoters that were specific only to the corresponding cell line, such as the core fibroblast markers BDKRB1[479] and COL1A1[480] only in IMR90 cells, the liver-specific genes MAT1A[481], MBL2[482], APOA2[483], and C8B[484] only in HepG2 cells, and highly expressed MCF-7 genes, PSMD6[485] and TRIM37[486], only in MCF-7 cells (**Supplementary Figure 8**).

Here, the genome-wide binding profile of TDG-N140A-CD to unmethylated active promoters in HEK293 cells apparently contradicts the experiments presented in **Figure 3B**, which demonstrated the high specificity of TDG-N140A-CD to the oxidized form of transiently transfected CMV-pCpGl plasmid. We sought to determine whether the same unmethylated CMV promoter – which, in this system, drives expression of TDG-N140A-CD and is integrated into the genome by lentivirus-mediated insertion – displays any TDG-N140A-CD binding within the same ChIP-seq data. To our surprise, stably integrated CMV promoter also accumulated significant TDG-N140A-CD peaks at the TSS in all three replicates (**Supplementary Figure 7H**) whereas the weaker SV40 promoter (also integrated as part of a selection marker cassette) did not contain any significant TDG-N140A-CD peaks, despite evident transcription from this promoter as observed by cell survival in the presence of the selection marker, blasticidin; this data supports selective binding of TDG to highly active promoters and demonstrates that newly introduced highly active promoters actively gain TDG binding.

Despite the ubiquitous presence of TDG in active gene promoters, knockout of TDG in HEK293 cells did not produce any significantly differentially oxidized CpGs genome-wide as measured by the Infinium MethylationEPIC BeadChip array with APOBEC-converted DNA. There were also no differences detected in HEK293 cell growth (**Supplementary Figure 9A**), overall RNA/DNA yield (**Supplementary Figure 9B-C**), or – as TDG was originally discovered as a glycosylase for repair of G:T mismatches – no apparent increased mutation burden of TDG knockout in 6 TDG-N140A-CD peaks

assessed by Sanger sequencing (**Supplementary Figure 9D**). Furthermore, of 28 genes with TDG-N140A-CD peaks in their promoters and 2 without – tested by RT-qPCR – there was no major effect of TDG knockout on expression, though one single statistically significant increase in gene expression was observed in *MYC* mRNA levels in TDG knockout cells (**Supplementary Figure 9E**).

TDG interacts with the MBD3/NuRD complex to bind unmethylated active promoters

The presence of TDG across active promoters regulated by RNA polymerase I, II, and III suggests a possible interaction between TDG and either a common factor of all transcription complexes (e.g., TBP or POL2RH) or a common recruitment of TDG by different transcription complex factors. To identify binding partners of TDG, we performed native co-immunoprecipitation of FLAG-tagged TDG-N140A-CD followed by identification of interacting proteins by LC-MS/MS, which identified 64 significantly enriched proteins in addition to TDG itself (**Supplementary Figure 10A, Supplementary File 3**).

Among top enriched proteins, we identified all components of a functional MBD3/NuRD complex[487]: MTA1, MTA2, p66-α/GATAD2A, p66-β/GATAD2B, HDAC2, RBBP4 and MBD3. MTA3 and HDAC1 – which could also participate in NuRD complexes in place of MTA1/2 and HDAC2 – were also enriched but were not statistically significant. As a whole, the NuRD complex was the highest enriched pathway among STRING-db local network clusters (CL:3874, FDR = 6.94×10^{-10}). Other proteins with epigenetic activity and/or known interactors of the NuRD complex were also detected: KDM1A, CXXC1, CSNK2A1, RBBP5, HELLS, DMAP1, SMARCA5, and HCFC1 among statistically significant proteins and SIN3A, OGT, and SET among proteins that did not reach statistical significance but were exclusively detected in TDG-N140A-CD samples. More broadly, the top enriched Reactome Pathways among the 64 hits were “chromatin

modifying enzymes" (HSA-3247509, FDR = 3.36 x 10^{-9}), "positive epigenetic regulation of rRNA expression" (HSA-5250913, FDR = 5.28 x 10^{-9}), and "epigenetic regulation of gene expression" (HSA-212165, FDR = 5.96 x 10^{-9}). Interestingly, both MBD3 (and the associated MBD3/NuRD complex) and SMARCA5 (which interact) have been implicated in the regulation of RNA polymerase I and III promoters[488-491], in addition to that of RNA polymerase II promoters, and MBD3/NuRD has been further implicated in the regulation of expression of genes containing oxidized cytosines[262,263] and is enriched at active promoters[492]. The interaction between TDG-N140A-CD and MBD3 was also confirmed by western blot and no interaction with MBD2 – which can form distinct NuRD complexes[493] with different functions – was detected (**Supplementary Figure 10B**). The interaction between TDG-N140A-CD and MBD3 – which is required for NuRD complex assembly[494] – was also observed by western blot in HepG2 and MCF-7 cell lines (**Supplementary Figure 10C**).

A physical interaction between MBD3/NuRD suggests that the proteins may co-localize in the genome. We therefore performed ChIP-seq of MBD3-bound DNA in HEK293, MCF-7, HepG2, and IMR90 cells for which we had performed anti-FLAG-TDG-N140A-CD ChIP-seq experiments. There was indeed considerable co-localization, wherein 53%, 51%, 33%, and 62% of MBD3 peaks were also sites of significant TDG-N140A-CD peaks in each cell line, respectively. For MCF-7 cells, we further assessed publicly available ChIP-seq data for additional members of the NuRD complex: MTA1, HDAC2, and GATAD2B. Here, we also found considerable co-localization, with 1,100 of 6,845 (MTA1), 2,825 of 14,787 (HDAC2), and 1,158 of 7,590 (GATAD2B) significant peaks of each protein also significantly bound by TDG-N140A-CD. Finally, all proteins (TDG, MBD3, MTA1, HDAC2, and GATAD2B) showed preferential binding to TSS of active genes with significant pol2 peaks in MCF-7 cells (**Supplementary Figure 10D**). To assess whether MBD3 is responsible for TDG-N140A-CD recruitment to unmethylated active promoters, we performed ChIP-qPCR of 3XFLAG-TDG-N140A-CD in control HEK293 cells and in MBD3 knockout cell lines from **Figure 4**. Knockout of MBD3

significantly reduced TDG-N140A-CD binding at *GAPDH*, *RNA5S*, and rDNA (*RNA45SN1*) promoters (**Supplementary Figure 11A-C),** suggesting that TDG-N140A-CD recruitment to its targets in HEK293 is partially mediated by its interaction with MBD3/NuRD.

We also assessed whether TDG could reciprocally recruit MBD3 to its genomic targets. ChIP-seq in HEK293 cells identified 3,056 significant MBD3 peaks. Half of these peaks (1,526) were also identified in cells with the N140A mutation in endogenous TDG, which had 5,032 significant MBD3 peaks in total. In TDGKO HEK293 cells, 1,306 of the 1,526 (86%) shared peaks were also significantly detected, suggesting that TDG is not required to recruit MBD3 to its binding sites. This is consistent with ChIP-qPCR experiments in **Figure 4** that demonstrated MBD3 binding to oxidized CMV-pCpGl in TDGKO cells, though in a different, oxidized transfected reporter context. Given that MBD3 has been shown to bind unmethylated active copies of the rDNA repeat promoter and its overexpression leads to demethylation of the rDNA promoter[490], we wondered if the TDG-MBD3 interaction may have a functional impact at the this locus, where we had reported significant FLAG-TDG-N140A0-CD binding in HEK293 cells (**Figure 5C**). We found that there were significant sites of FLAG-TDG-N140A binding in the rDNA repeat promoter also in HepG2 and IMR90 cells, though no binding was detected in MCF-7 cells (**Supplementary Figure 11D**). Furthermore, significant MBD3 peaks were detected in the rDNA locus in all cell lines, including HEK293 TDG-N140A and HEK293 TDGKO cell lines as well as public HEK293 MBD3 ChIP-seq data (**Supplementary Figure 11D**). Like MBD3[490], TDG was bound to significantly less methylated copies of rDNA (**Supplementary Figure 11E**), but the rDNA promoter was not oxidized in control HEK293 cells or TDGKO cells, nor was there any DNA methylation (from bisulfite-sequencing) differences between control and TDGKO cells (**Supplementary Figure 11F**). A minor effect of TDG knockout (hypermethylation) was only observed on a MBD3 knockout background, which, after correction for multiple testing, was only significant at a single CpG (**Supplementary Figure 11G**). Another major locus of shared significant

FLAG-TDG-N140A-CD and MBD3 peaks was the 5S rRNA gene cluster on chromosome 1: MBD3 peaks in this region were retained in TDGKO cells (**Supplementary Figure 11H**). Thus, all data suggest that TDG is not required for MBD3 recruitment to its targets.

Another top co-immunoprecipitated hit was SAFB1/2, two highly similar proteins that are known to interact with RNA polymerase II and are found in AT-rich scaffold/matrix attachment regions (S/MARs) which occur near actively transcribed genes: SAFB is thought to function as part of a "transcriptosome" complex which couples transcription initiation and RNA processing[495,496]. Accordingly, also enriched were a large number of proteins involved in the nuclear pore complex – KPNA3, NUP107, NUP93, NUP85, NUP62, RANGAP1 – and proteins involved in RNA processing – XRCC5, DDX21, HNRNPUL2, RALY, SRRT, NCBP1, ZNF326, SRSF9, DEK, RIOK1 – a phenomenon which has been previously observed for MBD3/NuRD complex member MTA1 and might similarly reflect a coordinated coupling of transcription and RNA processing[497]. This was also reflected in the fact that the STRING-db local network cluster "mRNA 3-end processing, and RNA recognition motif. (a.k.a. RRM, RBD, or RNP domain)" was the second highest enriched category (CL:1441, FDR = 3.98×10^{-8}) after "NuRD complex". Despite the strong overlap in binding occupancy between TDG-N140A-CD and TET1/TET2 and the existence of several shared binding partners (HDAC1, HDAC2, SIN3A, OGT, CXXC1[498]), no TET enzymes co-immunoprecipitated with TDG.

Binding sites of TDG-N140A-CD are not oxidized in HEK293 cells

Given our data that TDG-N140A-CD specifically binds oxidized CMV-pCpGI and can be used to differentiate oxidized CMV-pCpGI from unmethylated CMV-pCpGI in transfected mixtures, we anticipated that TDG-bound regions in ChIP-seq data would be enriched for naturally oxidized CpGs. We performed genome-wide APOBEC-seq of total HEK293 DNA (input) and TDG-bound immunoprecipitated DNA (FLAG-IP). We note that while

APOBEC-seq does not directly discriminate 5hmC from 5fC and 5caC, it reveals, in this context, whether TDG-N140A-CD binding sites contain cytosines exhibiting any of the oxidized states. To improve mapping, we first used the HEK293 ChIP-seq data to identify short sequence differences (SNPs and INDELs) between HEK293 cells and the human reference genome and generate a “HEK293 genome” that was then used for APOBEC-seq alignment. Genome-wide APOBEC-seq achieved a mean ± SD conversion rate of 99.62 ± 0.19%, which is comparable to the rates achieved by bisulfite sequencing kits[31,499], and >10X coverage of approximately 48% of genomic CpGs on average.

We were surprised to find that TDG-bound DNA was uniformly not oxidized, with CpG oxidation rates across three replicates of 0.8%, 0.5%, and 0.6% just above estimated APOBEC conversion failure rates (0.29%, 0.20%, and 0.22%, respectively) with a complete absence of highly oxidized sites (**Supplementary Figure 12**) and oxidation levels decreasing further towards TDG-N140A-CD peaks centers (**Figure 6A**). Interestingly, despite the fact that input DNA oxidation also decreased towards TDG-N140A-CD peak centers, there was a slight increase in oxidation at the centers compared to their edges (**Figure 6B**) together suggesting that, in a population of cells, oxidation can increase in the regions that can be bound by TDG, but TDG-N140A-CD is bound to these regions when they are less oxidized. Any detectable oxidation decreased sharply towards background levels in TSS (**Figure 6C-D**)**.**

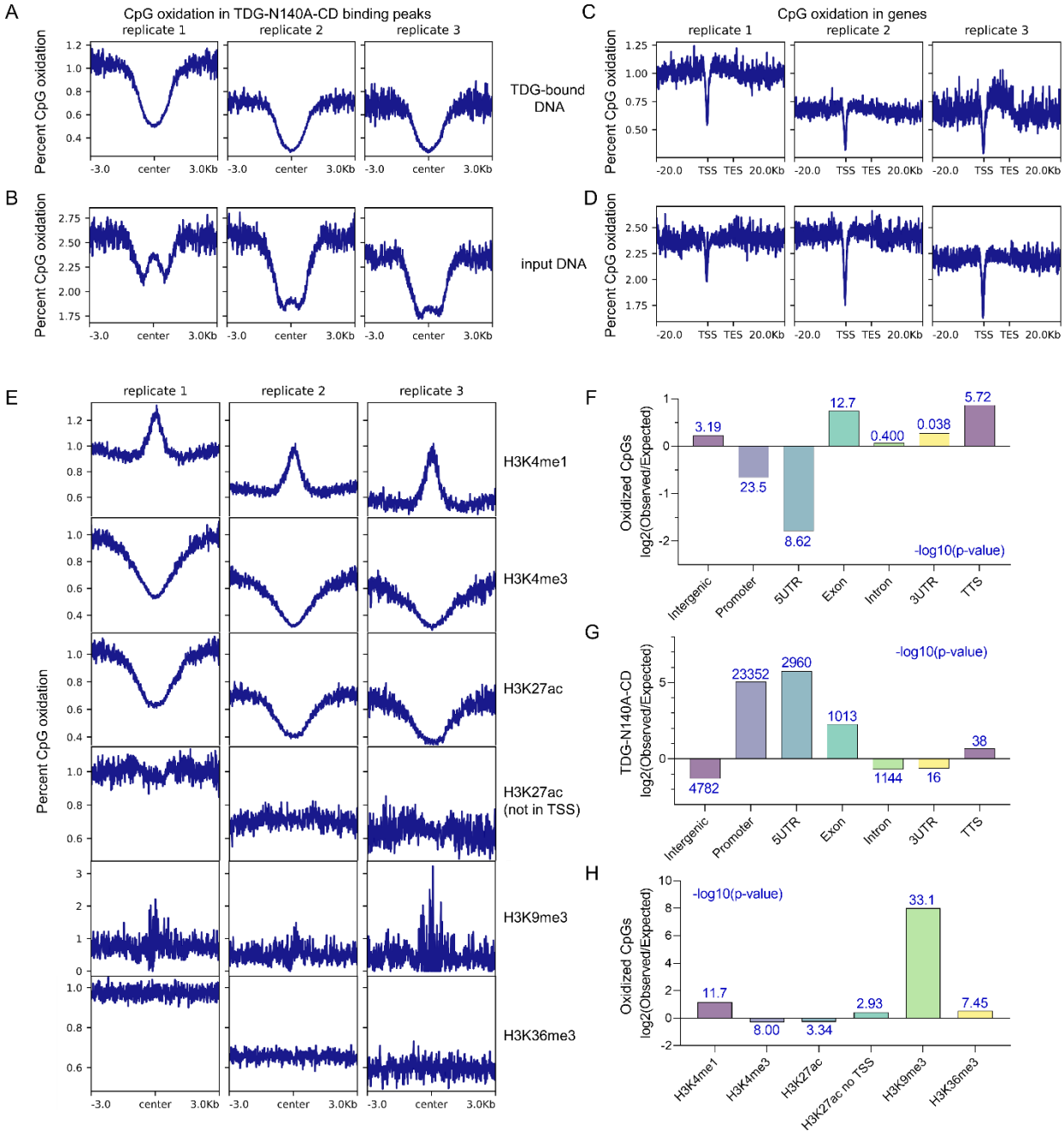

A
CpG oxidation in TDG-N140A-CD binding peaks
replicate 1
replicate 2
replicate 3
Percent CpG oxidation
TDG-bound DNA
B
input DNA
C
CpG oxidation in genes
D
-3.0
center
3.0Kb
-20.0
TSS
TES
20.0Kb
E
H3K4me1
H3K4me3
H3K27ac
H3K27ac (not in TSS)
H3K9me3
H3K36me3
F
Oxidized CpGs
log2(Observed/Expected)
-log10(p-value)
3.19
23.5
8.62
12.7
0.400
0.038
5.72
Intergenic
Promoter
5UTR
Exon
Intron
3UTR
TTS
G
TDG-N140A-CD
4782
23352
2960
1013
1144
16
38
H
11.7
8.00
3.34
2.93
33.1
7.45
H3K4me1
H3K4me3
H3K27ac
H3K27ac no TSS
H3K9me3
H3K36me3

Figure 6. Genome-wide APOBEC-seq in HEK293 cells. (A-B) Percent CpG oxidation averaged over all significant 3XFLAG-TDG-N140A peaks from ChIP-seq data, plotted as CpG distance from the center of the peak, for CpGs in APOBEC-seq data with a minimum of 10X coverage in 3 biological replicates of ChIP DNA bound to 3XFLAG-TDG-N140A (A) or corresponding input DNA (B). (C-D) Percent CpG oxidation averaged over gene structures, plotted as a function of distance from the transcription start site (TSS) or transcription end site (TES), for CpGs in APOBEC-seq data with a minimum of 10X coverage in 3 replicates of ChIP DNA bound to 3XFLAG-TDG-N140A (C) or corresponding input DNA (D). The region between TSS and TES is scaled to 10 kb. (E) Percent CpG oxidation from APOBEC-seq, filtered for >10X coverage, in 3 biological replicates of input DNA samples, plotted as in (B) but as a function of distance relative to centers of significant peaks of H3K4me1, H3K4me3, H3K27ac, H3K27ac outside of promoters, H3K9me3, and H3K36me3 binding from ENCODE data. (F) Log2-scaled observed/expected ratios of oxidized CpG occurrence in each indicated genomic element. Blue values above bars indicate negative log10 p-values associated with the observed/expected ratio, calculated by Fisher's exact test. Expected values were calculated based on all sequenced CpGs (regardless of oxidation state) that passed indicated thresholds. (F) Log2-scaled observed/expected ratios of significant TDG binding peaks in genomic elements. (G) Same as (F) but for significant peaks of each indicated histone mark. For reference, -log10(0.05) is approximately equal to 1.3.

Using public data from ENCODE[500], we assessed the relation between CpG oxidation and histone marks. In accordance with previous data, CpG oxidation peaked in poised enhancer regions marked by H3K4me1 but also in H3K9me3-bound regions (**Figure 6E**). In contrast, CpG oxidation levels declined in regions with the active marks H3K4me3 and H3K27ac. However, no such decline was observed in H3K27ac regions when they were not in TSS. There were no changes in CpG oxidation levels in H3K36me3 peaks. Furthermore, there were no significantly hyper-oxidized CpGs in TDG-bound DNA compared to input DNA, suggesting that either enrichment of oxidized DNA by TDG-N140A-CD by ChIP is a phenomenon specific to transfected plasmid DNA or that there is little to no CpG oxidation in HEK293 cells. In an effort to identify any consistently oxidized CpGs while reducing multiple testing burden, we called oxidized sites as those with >20% oxidation across all three replicates of input and/or FLAG-IP samples, supported by >20X coverage, and an absence of any HEK293-specific mutations within 150 bp in either direction. For these thresholds, we disregarded one FLAG-IP replicate due to a ~50% reduced sequencing depth compared to other samples. 568 CpGs in FLAG-IP samples and 1,889 CpGs in input samples passed these filters, 212 of which were the same CpGs. Of the 356 FLAG-IP specific oxidized CpGs, 318 may not be FLAG-IP specific because they did not pass the coverage threshold in at least one input sample. The remaining 38 CpGs were not significantly differentially oxidized between input and FLAG-IP samples after correction for multiple testing. We therefore conclude that there are no highly oxidized CpGs specifically enriched by ChIP against FLAG-tagged TDG-N140A-CD. Reciprocally, 328 (17%) of oxidized CpGs detected in input samples were not sufficiently covered in at least one FLAG-IP sample and were discarded from the following analysis. Of the 1,561 highly oxidized CpGs in input samples that were covered in FLAG-IP samples, 1,542 were less oxidized in FLAG-IP samples, though only one CpG reached statistical significance after correction for multiple testing: this CpG (chr2:130341919) is located in a TDG-N140A-CD peak and pol2 peak near the TSS of the *CCDC115* gene, supporting the

notion that TDG-N140A-CD is bound to DNA copies that are less oxidized (as seen in **Figure 6A,B**).

To avoid regional biases as a function of TDG binding, we profiled the genomic locations of the highly oxidized CpGs only in input samples. As expected, promoters were statistically depleted of oxidized CpGs, as well as nearby 5' UTRs (**Figure 6F**) which is where TDG is most enriched (**Figure 6G**). Accordingly, oxidized CpGs were significantly enriched in H3K4me1- and H3K9me3-marked regions as well as regions marked with H3K27ac that were not in promoters and significantly depleted from regions marked with H3K4me3 and H3K27ac (**Figure 6H**). Though increased oxidation in H3K4me1-bound regions is consistent with the literature, oxidation in regions marked by H3K9me3 is previously unreported and may have important implications in relation to recent observations of the involvement of H3K9me3 in tissue-specific gene expression and cell identity[501].

Oxidized genes in the adult mouse brain cortex are highly tissue-specific and MBD3 and TDG are depleted from these regions

As HEK293 cells have nearly undetectable of levels of oxidized cytosines[233], exacerbated by a failure of their enrichment by 3XFLAG-TDG-N140A, we sought to study oxidation *in vivo*. Though APOBEC-seq does not distinguish between oxidized states, it provides a broader profile of oxidation that may reveal insight into regions that are regulated by the active DNA demethylation pathway and allows us to identify how MBD3 and TDG regulate this pathway in these regions. Given that the brain is the adult tissue with the highest cytosine oxidation rates[233], we applied genome-wide APOBEC-seq to the cortices of three 8-week-old adult female mice. Numerous studies which have previously mapped oxidized cytosines (5hmC[248,250,251,253,502,503], 5fC[244,246,251,257-259], and 5caC[244,246,247]) using other methods have consistently observed their enrichment in poised (H3K4me1-marked) and active enhancers (H3K4me1- and H3K27ac-marked),

though these studies report contradictory findings about their presence at active transcription start sites. To further elucidate oxidized cytosine distribution with APOBEC-seq, we generated ChIP-seq profiles of H3K4me1, H3K27ac, and the initiating form of RNA polymerase II (phosphorylated on serine 5, pol2-PS5) (**Supplementary Figure 13**) as well as APOBEC-seq of each ChIP DNA for each cortex, in addition to APOBEC-seq of total (input) DNA. APOBEC-seq summary statistics – including coverage and conversion rates – are available in **Supplementary Table 1**.

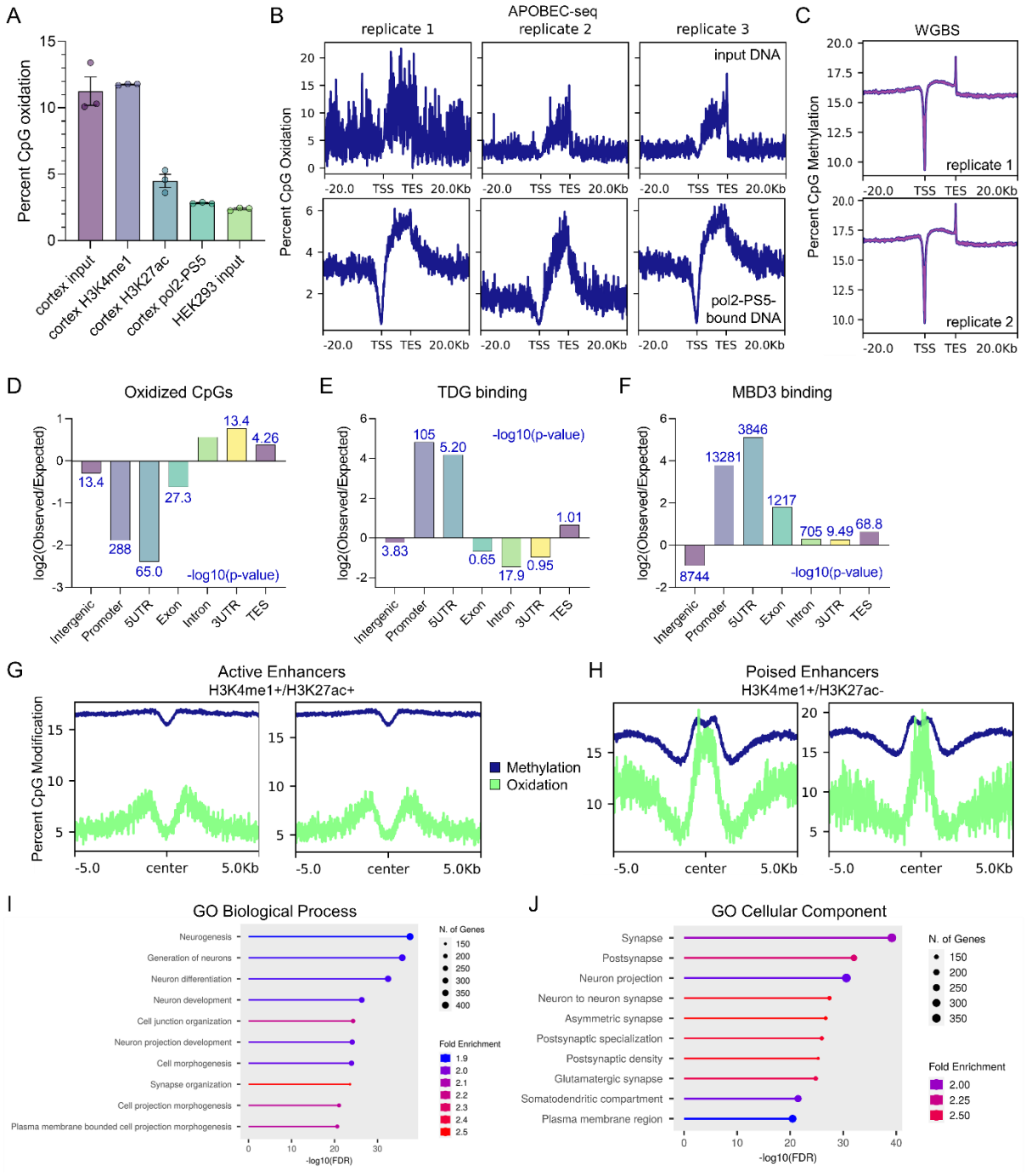

A
Percent CpG oxidation
cortex input
cortex H3K4me1
cortex H3K27ac
cortex pol2-PS5
HEK293 input
B
APOBEC-seq
replicate 1
replicate 2
replicate 3
input DNA
pol2-PS5-bound DNA
Percent CpG Oxidation
-20.0 TSS TES 20.0Kb
C
WGBS
Percent CpG Methylation
replicate 1
replicate 2
D
Oxidized CpGs
log2(Observed/Expected)
-log10(p-value)
Intergenic
Promoter
5UTR
Exon
Intron
3UTR
TES
E
TDG binding
F
MBD3 binding
G
Active Enhancers
H3K4me1+/H3K27ac+
Percent CpG Modification
Methylation
Oxidation
-5.0 center 5.0Kb
H
Poised Enhancers
H3K4me1+/H3K27ac-
I
GO Biological Process
Neurogenesis
Generation of neurons
Neuron differentiation
Neuron development
Cell junction organization
Neuron projection development
Cell morphogenesis
Synapse organization
Cell projection morphogenesis
Plasma membrane bounded cell projection morphogenesis
-log10(FDR)
N. of Genes
Fold Enrichment
J
GO Cellular Component
Synapse
Postsynapse
Neuron projection
Neuron to neuron synapse
Asymmetric synapse
Postsynaptic specialization
Postsynaptic density
Glutamatergic synapse
Somatodendritic compartment
Plasma membrane region

Figure 7. CpG oxidation in the adult mouse cortex. (A) Average percent CpG oxidation across all aligned APOBEC-seq CpG calls from input, H3K4me1-ChIP, H3K27ac-ChIP, and pol2-PS5-ChIP DNAs, compared to that of input in HEK293 cells. (B) Percent CpG oxidation averaged over gene structures, plotted as a function of distance from the transcription start site (TSS) or transcription end site (TES), for CpGs in APOBEC-seq data with a minimum of 10X coverage in 3 cortices for input DNA (top 3 panels) or ChIP DNA bound to pol2-PS5 (bottom 3 panels). The region between TSS and TES is scaled to 10 kb. (C) Same as (B) but public WGBS data of total DNA from two replicates of mouse cortex. (D) Log2-scale observed/expected counts for oxidized CpGs in each indicated genetic element. Blue values above bars indicate negative log10 p-values associated with the observed/expected ratio, calculated by Fisher's exact test. No -log10(p-value) is depicted for introns because Fisher's exact test returned a p-value of 0 and thus the enrichment should be considered highly significant. (E-F) Log2-scale observed/expected counts for significant binding peaks of TDG (E) and MBD3 (F) in each indicated genetic element in mouse cortex samples. Blue values above bars indicate negative log10 p-values associated with the observed/expected ratio, calculated with homer as described in the methods. (G-H) Average percent CpG methylation (blue) and CpG oxidation (green) in APOBEC-seq data of mouse cortexes, plotted over peaks of significant H3K4me1 and H3K27ac binding (G, active enhancers) or peaks of significant H3K4me1 binding without significant H3K27ac binding (H, poised enhancers), centered at peak centers and ±5 kb on either side. For reference, -log10(0.05) is approximately equal to 1.3. (I-H) Gene ontology term enrichment analysis depicting terms significantly enriched within GO Biological Process (I) or GO Cellular Component (J) pathways among genes with high-confidence oxidized CpGs.

Oxidation rates in mouse cortex were indeed higher than in HEK293 cells (**Figure 7A**). Oxidation rates of specific CpGs in any one cortex were significantly predictive of oxidation rates of the same CpG in the other samples (**Supplementary Figure 14A-C**) and the maximum oxidation of any CpG that was covered at least 10X in all three input samples (and not near identified SNPs or INDELs) was 33.2% (27.3%, 41.7%, 30.8% in the 3 samples; chr2:48000475 in Gm13481, a non-coding RNA); however, the average percent oxidation of all ≥10X covered CpGs was 3.778 and the median was 2.944 (**Supplementary Figure 14D**). In the context of gene structures, CpG oxidation occurred primarily in gene bodies (**Figure 7B**), with pol2-bound genes characterized by a sharp decrease in CpG oxidation at the TSS and a steady increase in oxidation towards a maximum at the TES (**Figure 7B**), which was not seen in HEK293 APOBEC-seq data, perhaps reflecting the overall lack of oxidation in this cell line (**Figure 6D**). This pattern is reminiscent but not identical to methylation (as measured by whole-genome bisulfite sequencing (WGBS) of mouse cortices, which also detects 5hmC) which decreases towards the TSS and shows a sharper increase in the TES (**Figure 7C**). In contrast to methylation, which is elevated in the gene bodies of active genes (**Supplementary Figure 15A**), oxidation levels are reduced in the bodies of pol2-bound genes (**Supplementary Figure 15B**). Oxidized cytosines (>5% in all input samples) were significantly depleted from intergenic regions and from promoters and 5' UTRs but enriched in introns, 3' UTRs, and TES (**Figure 7D**). As in HEK293 cells, this oxidation pattern is strikingly inverse to that of TDG binding assayed by ChIP-seq in the same cortices, which is significantly enriched in promoters and 5' UTRS and depleted from introns (**Figure 7E**), suggesting that active DNA demethylation by TDG may sculpt CpG oxidation levels at these regions.

We detected 127 significant peaks of endogenous TDG binding that were replicated in at least two of three cortices. 103 (81%) of TDG peaks were also significant peaks of MBD3 binding, including all identified TDG peaks in promoters and 5' UTRS. Interestingly, 18 of these shared peaks were in RNA4.5S genes, located primarily in a

cluster on chromosome 6 (**Supplementary Figure 15C**). While these shared peaks were fully devoid of CpG oxidation, oxidized CpGs were nonetheless significantly enriched in rRNA genes ($p = 6.590733 \times 10^{-62}$), a disparity explained by the presence of oxidized CpGs in the regions immediately flanking these peaks (**Supplementary Figure 15D**) and suggesting that, here, TDG also maintains low oxidation at its binding sites. This parallels the findings in HEK293 cells, where FLAG-TDG-N140A-CD and MBD3 were both enriched at related RNA5S genes (**Figure 5C**, **Supplementary Figure 11**).

Like TDG, MBD3 peaks were also primarily found in promoters and 5' UTRs, but also in exons, and to a lesser extent, introns 3' UTRs and TES (**Figure 7F**).
Similar to DNA methylation (WGBS), oxidation decreases towards the peak centers in DNA bound to pol2-PS5 and H3K27ac but increases towards H3K4me1 peak centers (**Supplementary Figure 15E**). Interestingly, in active enhancers (regions outside of promoters marked by H3K27ac) while methylation decreased, a distinct pattern of oxidation was observed: oxidation sharply increased at the edges of the peak and decreased again towards the peak center (**Figure 7G**), which may be indicative of a first wave of broader TET-mediated oxidation followed by a narrower TDG activity as enhancers become active. In poised enhancers (marked by H3K4me1, without H3K27ac), methylation increased ~5% from its minimum at the edges to its maximum near the peak center, with a small ~1% decrease at the peak center. In contrast, oxidation exhibited a larger ~10% increase and did not decrease at the peak center (**Figure 7H**).

We then defined 7,748 high-confidence oxidized CpGs as those supported by at least two oxidized CpG reads per input replicate and replicated in at least two of the three cortex samples. To again exclude the potential influence of sequence variants on oxidation calling, this list was then pruned for CpGs that overlapped a leniently filtered list of SNPs or INDELs identified from ChIP-seq data (leaving 7,706) and pruned further for those that were greater than 50 bp away (leaving 6,714, **Supplementary File 4**). We

selected three such CpGs for validation in three additional independent cortex samples along with deeper sequencing by targeted bisulfite- and APOBEC-pyrosequencing, which revealed methylation between 43% and 86% (**Supplementary Figure 15F**) and confirmed CpG oxidation in this independent group of animals. CpG oxidation ranged from 16% to 43% across the three CpGs and percent oxidation at each CpG was nearly identical between the three cortices (**Supplementary Figure 15G**).

Using a dataset of tissue-specific enhancers across 22 mouse tissues[504], we found that cortex-specific enhancers were the most enriched ($p = 2.20 \times 10^{-15}$) tissue-specific enhancer group among all tissue-specific enhancers with high-confidence oxidized CpGs, while enhancers specific to most other tissue-types were depleted (**Supplementary Figure 15H,I**). To further demonstrate the marked tissue-specificity of enhancer CpG oxidation, we used an independent study which linked enhancers to the genes regulated by these enhancers[505] and found that genes linked to enhancers containing oxidized CpGs were over-represented in a list of gene ontology categories that were almost exclusively neuron-specific, such as dendrite development, neurogenesis, and neuron differentiation (GO Biological Processes) and axons, synapses, neuron-to-neuron synapses (GO Cellular Component) (**Supplementary Figure 15J,K**). Finally, genes which contained oxidized CpGs (wherein oxidation primarily occurred in gene bodies (**Supplementary Figure 16**)) were likewise enriched for exclusively neuron-specific gene ontology categories (**Figure 7I,J**) to a greater degree than methylated CpGs defined in the same manner (**Supplementary Figure 17**)

14.4 Discussion

We show here that APOBEC-seq is an efficient bisulfite-free base-resolution sequencing technique to directly detect oxidized cytosines. APOBEC3A conversion has been demonstrated to be efficient and nondestructive in EM-seq[30] and ACE-seq[473] protocols, which prioritize detection of 5mC and 5hmC, respectively, through additional enzymatic steps. While both our data and NEB data (see Supplementary Figure 3C in

Vaisvila et al.[30]) demonstrate that 5caC is completely resistant to deamination by APOBEC3A, Vaisvila et al. also demonstrated complete resistance to deamination of 5fC and an approximately 50% resistance of 5hmC while we report only limited resistance of 5fC and 5hmC to deamination. The source of this discrepancy is unclear but may stem from differences in oligonucleotide purities or in detection methods (LC-MS vs. sequencing). Until this is resolved, the magnitude to which APOBEC-seq preferentially detects 5caC over 5fC/5hmC remains unclear. While this has implications in developing conclusions from genome-wide APOBEC-seq data, it does not minimize the utility of APOBEC-seq in the analysis of the dynamics of active DNA demethylation of oxidized 5caC transfected promoter-reporter plasmids, which are oxidized fully to 5caC by an *in vitro* TET reaction.

What is the role that 5-methylcytosine oxidation plays in transcription and DNA demethylation? APOBEC-seq, in combination with several genetic perturbations, plasmid variations, and ChIP experiments, have allowed us to comprehensively investigate these processes, and notably, in cells that are both human in origin and differentiated. We mainly used transient transfection of *in vitro* methylated and oxidized reporter into HEK293 cells since this allowed us to focus on a specific oxidation form (5caC) and to study the causal interrelationship between defined states of oxidation, transcription and active demethylation. Altering oxidation of endogenous genes in cells ex vivo or tissues in vivo involves using TET enzymes and since these enzymes have additional transcriptional activities which are methylation independent[2], it is impossible to dissociate the effects of oxidation from other activities of TET. In addition, any study in living cells of oxidized bases is confounded by passive demethylation driven by DNA replication. Using this model we first show that oxidized cytosines are rapidly lost in HEK293 cells and second that this demethylation is accompanied by increased expression, which is dependent on endogenous activity of TDG as has been previously shown in mouse embryonic stem cells (mESCs)[506] without assessment of DNA (de)methylation states. In absence of TDG, both demethylation and transcription are

inhibited, and the reporter remains poorly expressed and oxidized supporting a causal role for active demethylation in transcriptional activation of the oxidized reporter and ruling out the possibility that 5caC is an independent epigenetic signal for gene activation. The small activity of the oxidized promoter in TDGKO cells is a possible reflection of the residual demethylation of oxidized promoters in these cells. This minor demethylation observed in TDGKO cells might be driven by alternative pathways of removal and replacement of the oxidized base by the NEIL family of DNA glycosylases[506]. We report that oxidized reporter reactivation is common to multiple human cell lines and that TDG mediates this effect, leading to RNA polymerase II recruitment as determined by ChIP assays with either RNApolII or RNApolII PS5 antibodies. Similarly, the TATA binding protein TBP is not bound to either methylated or oxidized promoters.

We further report a trans-activation capacity of TDG on methylated and unmethylated DNA that is independent of its DNA demethylation capacity which suggests that transcriptional activation and demethylation are partially decoupled; it is therefore unclear if demethylation is required for transcription or vice versa. To address this question, we first show that transcription inhibitors reduce demethylation to a certain extent but do not completely inhibit demethylation. Second, a transcriptionally deficient oxidized plasmid is still demethylated, demonstrating that transcription is not required for TDG-mediated active DNA demethylation. Demethylation is seen also in weak promoters such as SV40 or TATA-less CMV promoter, supporting the idea that demethylation is not dependent on transcription activity. Third, to determine whether transcription activation is caused by TDG triggered demethylation or by its DNA methylation independent transactivation activity, we knock out APEX1 to dissociate excision repair – which is required for replacement of the oxidized bases by cytosine – from transcription. However, to our surprise and despite previous reports that APEX1 is responsible for all base excision repair in HEK 293 cells[478], there is only a small inhibition of demethylation by APEX1 removal, which corresponds to a limited inhibition

of expression. Although these data possibly challenge our understanding of the sequential enzymatic steps leading to demethylation of oxidized cytosines triggered by TDG, it is possible that an alternative pathway catalyzed by NEIL proteins substitutes for APEX1[506]. This question remains to be addressed. Full evidence that demethylation triggered by TDG triggers transcription would require complete dissociation of TDG transcription activation from replacement of oxidized cytosines by downstream excision and repair.

Although our results demonstrate that transcription is not necessary for triggering demethylation by TDG, we noted partial reduction in demethylation when transcription is inhibited, for example when we used a single oxidized CG promoter (**Supplementary Figure 5K)**. Similarly, although oxidized CG dense sequence is not necessary for demethylation it its extent is reduced relative a full dense oxidized CG promoter. These suggest that transcription as well as CG density might be contributing to the TDG-triggered demethylation reaction by yet unknown mechanisms.

Though oxidized reporter expression in TDGKO cell lines remains consistently above that of methylated reporter (possibly corresponding to residual demethylation in these cells potentially mediated by the NEIL family of DNA glycosylases[506]) we report that the catalytic N140A mutation in TDG suppresses expression further than TDGKO. This is likely due to an exceptionally strong binding activity of mutant TDG to oxidized DNA, which can be used to purify oxidized DNA from transfected mixtures of unmethylated and oxidized reporter plasmid. If recombinant mutant TDG performs similarly in an in vitro assay, this may prove to be a highly efficient method for the enrichment of oxidized DNA as a means for higher-resolution sequencing (e.g. with APOBEC-seq) or diagnosis in diseases where 5caC is a biomarker[507], in a way that parallels the advantages of purifying methylated DNA by MBD proteins rather than anti-5mC antibodies[508].

Since oxidized CpG-containing promoters are silenced, we tested whether MBDs are involved in silencing oxidized promoters. We show that the MBD family of proteins bind transfected oxidized plasmid DNA – with a particularly specific binding by MBD3 – though there is little functional effect of knockout of any individual MBD protein on reporter expression and demethylation. Our data suggest that each of the MBD proteins are not critical for silencing of the oxidized promoter in our cells although there might be redundancy of binding of MBDs that was not investigated here. On the contrary, MBD2 and MBD1 knockout result in reduced activity of oxidized promoters which is consistent with a role in gene activation. The GR domain of MBD2 might have some suppressive effect since its removal increases expression of the oxidized promoter in TDGKO cells. However, while all MBDs showed binding activity to oxidized promoters, competition assays with TDG-N140A-CD suggest distinct interactions with TDG. While MBD1 and MECP2 compete with TDG for binding to oxidized CG promoters and might act as steric inhibitors of TDG triggered demethylation (which is consistent with the increased demethylation in MBD1 KO), MBD3 binding is not competed with by TDG-N140A-CD. Though it would be interesting to compare the binding affinities of the MBD proteins to oxidized DNA, in this assay, possible differences in antibody qualities/affinities prevent comparisons of the magnitude of binding across different proteins. We find that the binding profiles we determined for each MBD family member to be strikingly parallel to those determined in a recent *in vitro* study which profiled the binding properties of recombinant MBD family members to all oxidized 5mC derivative combinations using electromobility shift assays (EMSAs)[509]. Though most experiments on oxidized plasmid dynamics in this study were conducted in HEK293 cells, there appears to be cell-type dependent variability in the expression of oxidized plasmids; future experiments that compare different cell types could aim to resolve the molecular basis for this differential activity. We noted that demethylation rates of oxidized CMV-pCpGl varied across the same time point in different experiments. This is likely due to combination of variable transfection efficacy and a more major effect of washing efficacy, as any plasmids that did not enter the nucleus would not be demethylated and thus contaminate the signal.

We have found a two-step lysis of cell membrane followed by nuclear envelope, as is common in ChIP protocols, to be critical to maximize the detection of demethylation in transfected oxidized plasmids. For this reason, we only compare demethylation rates within individual experiments.

To further understand the role of oxidized CGs and TDG in regulating transcription we examined the landscape of TDG binding and oxidized cytosines in both human cell lines and mouse cortices. ChIP-seq experiments in this study demonstrated a ubiquitous binding activity of TDG to active unmethylated and unoxidized promoters independent of the RNA polymerase type (I, II, or III) that drives promoter expression. The fact that TET and TDG are ubiquitously present across unmethylated active promoters is consistent with the idea that active DNA demethylation may safeguard promoters from aberrant hypermethylation and gene silencing. The widespread presence of TDG at active TSS across the genome and its persistence in differentiated cells and tissues and not just embryonic cells is consistent with an important role in safeguarding transcription state and epigenetic integrity. TDG is bound to completely unmethylated and unoxidized sequences; the residence of epigenetic factors and enzymes at regulatory regions even though their natural substrate has long been acted upon is unexpected but might reflect a mechanism to ascertain the fidelity and prevent even a small drift in the epigenetic landscape. The positioning of TDG at active TSS while depleting it from other sequences may allow oxidation to accumulate in other genomic areas, thus maintaining an oxidation landscape in nondividing tissues such as the brain (**Figure 7**). This might explain why oxidized CpGs, which are typically intermediates in a demethylation reaction, are more stable in the brain. In diving cells, any spurious oxidation is poorly maintained during cell division resulting in a low level of oxidized CpGs across the genome. While protection from spurious methylation-oxidation might be one of the roles that TDG plays, it might also protect transcription start site from deamination, thus protecting the genomic integrity of promoters. This binding of TDG at active promoters may explain why promoters which are efficiently methylated by tools such as dCas9-

DNMT3A fusion proteins are then rapidly demethylated upon termination of dCas9-DNMT3A expression[310,510,511], a persistent problem in the DNA methylation editing field and one that has only been resolved by the combinatorial targeting of multiple epigenetic repressors. Our experiments demonstrating that methylation of stably integrated methylated CMV promoter is better maintained in TDG knockout cells provide further evidence for this hypothesis. Given the ubiquitous binding of TDG to active promoters, we were surprised to find that TDG knockout had minimal effects on CpG oxidation and HEK293 cell biology, but this could be explained by low TET activity[512] and consequently low oxidation levels[233] in HEK293 cells; a lack of effect of TDG knockout on gene expression has also been independently reported[513]. An alternative hypothesis is that TDG plays a different role in promoters that is independent of DNA methylation/oxidation.

The source of the discrepancy between the specific binding of 3XFLAG-TDG-N140A-CD to oxidized transfected CMV promoter compared to its ubiquitous binding of unmethylated active endogenous promoters remains unclear. There may be a difference in the cellular response to ectopic plasmid DNA as opposed to genomic DNA or binding of TDG-N140A-CD to unmethylated active promoters may reflect a physiologically significant but weaker binding activity than to oxidized DNA. Indeed, endogenous TDG binding to both unmethylated and methylated active CMV-pCpGl – though much lower than to oxidized CMV-pCpGl – was above background levels (TDG knockout) (**Figure 2C**). This also suggests that TDG-N140A-CD may be recruited to unmethylated active promoters by interactions with other proteins. We used co-immunoprecipitation followed by LC-MS/MS to identify potential interactors of 3XFLAG-TDG-N140A-CD and identified all components of an MBD3/NuRD complex. We used MBD3 ChIP-seq in numerous cell lines and *in vivo* in the mouse cortex and report that MBD3 similarly binds active promoters, and its knockout reduces TDG binding to its targets. The distinct binding of MBD3/NuRD in the promoters of active genes is contradictory to is classical depiction as a repressive complex[493], which is largely

assumed from the nature of its components. However, in addition to the results presented in this work, preferential binding to active unmethylated TSS has been previously demonstrated for MBD3[492,514] and for the classically repressive NuRD complex members HDAC1/2[515]. For the latter proteins, increased binding to TSS correlates with increased transcription and RNA polymerase II binding and additional work has demonstrated that they are required for transcription of core regulatory genes[516]. These data, in addition to our experiments which demonstrate that MBD3/NuRD recruits TDG to active promoters, suggest the need for a re-evaluation of the role of the MBD3/NuRD complex in gene expression. The interaction between MBD3/NuRD and TDG may prove to be the missing mechanistic link that explains DNA demethylation of promoters upon MBD3 overexpression[517] and thus explain the requirement of MBD3 for early embryonic development[518], cell pluripotency[494], and pluripotent cell development[519], by mediating the TDG-dependent active DNA demethylation that appears to be required for these processes[520,521].

Finally, we demonstrate the technical feasibility of genome-wide APOBEC-seq in mouse cortices and find that oxidized CpGs are found in highly tissue-specific sets of genes and enhancers, which has been reported previously[248,522]. We find that oxidation in these regions may be maintained by the absence of TDG binding. The tissue specificity of this oxidation suggests an important role. However, our transient reporter assays suggest that oxidized CpGs are silencing gene expression as much as methylated CpGs and are rapidly and actively removed. These sites remain oxidized on in the absence of TDG activity and it remains unclear how they contribute to gene regulation. The fact that oxidation is tissue-specific and percent oxidation rates are nearly identical across the same CpG of different animals suggests that oxidation, which can only be 0%, 50% or 100% for any single double-stranded DNA molecule, is likely to be further specific for cortical cell subtypes. It may therefore be important in the future to study oxidation signatures in different cells of the brain using single-cell APOBEC sequencing.

14.5 Methods

Plasmids

CMV-pCpGl plasmid was generated by amplifying the CMV promoter/enhancer sequence from pMD2.G (Addgene, plasmid no. 12259) with primers that added a 5' BamHI site and a 3' HindIII site (all primers in **Supplementary File 5**) using standard Taq polymerase (Thermo Fisher, cat. no. EP0401) following the manufacturer's protocol and 10X Taq Buffer with KCl. PCR products were run on a 1% agarose gel, stained with ethidium bromide, and purified from the agarose gel with the QIAEX II Gel Extraction Kit (QIAGEN, cat. no. 20021). Purified PCR products were ligated into pCR®4-TOPO TA vector (Thermo Fisher, cat. no. 450030) following manufacturer's instructions with a 30-minute room temperature ligation step and transformed into NEB® Stable Competent E. coli (NEB, cat. no. C3040H) by heat-shock transformation. *E. coli* were plated onto LB-agar dishes containing 100 µg/ml carbenicillin, grown overnight at 37 °C, inoculated into liquid LB containing 100 µg/ml carbenicillin and grown overnight in a 37 °C shaking incubator. Plasmid DNA was purified from *E. coli* using the Presto™ Mini Plasmid Kit (Geneaid, cat. no. PDH300) and assessed by Sanger sequencing with the M13R primer to verify the sequence of the amplified CMV promoter. Sequence-verified plasmid as well as pCpGl vector[299] were digested with BamHI-HF and HindIII-HF (NEB, cat. nos. R3136S, R3104S) and purified by gel extraction as described above. Purified plasmids were ligated overnight at 16 °C with T4 DNA ligase (NEB, cat. no. M0202S) according to manufacturer's protocol, heat inactivated at 65 °C for 10 min and transformed into One Shot™ PIR1 Chemically Competent E. coli (Thermo Fisher, cat. no. C101010). Resulting plasmids were verified by restriction digest and purified with the Plasmid Maxi Kit (QIAGEN, cat. no. 12163). TATA box mutation of CMV-pCpGl was achieved by a TATATA -> TGTTGC mutation using primers listed in **Supplementary File 5** and the Q5® Site-Directed Mutagenesis Kit (NEB, cat. no. E0554S) according to the manufacturer's protocol. The pCpGl plasmid with a single CpG site was constructed by annealing of two oligonucleotides (**Supplementary File 5**) by resuspension of each to

100 µM in annealing buffer (10 mM Tris, pH 8.0, 50 mM NaCl, 1 mM EDTA), mixing 50 µL each, heating to 99 °C in a thermocycler and ramping the temperature down to 25 °C over 45 min. The duplexed DNA was then purified with the Monarch® PCR & DNA Cleanup Kit (NEB, cat. no. T1030L) according to the manufacturer's protocol, resuspended in PCR-grade water, and processed with BamHI and HindIII for ligation into pCpGl as described above. The pCpGl plasmid encoding the single CpG and a CpG-less CMV promoter was constructed by gene synthesis (Integrated DNA Technologies) (**Supplementary File 5**) and resuspended to 100 ng/µL in PCR-grade water which was processed with BamHI and HindIII for ligation into pCpGl as described above. The SV40-pCpGl plasmid was constructed the same way as CMV-pCpGl except that the SV40 promoter was amplified from pLenti-gRNA-puro (Addgene, plasmid no. 180426) and the 5' cloning site was SpeI (**Supplementary File 5**) and SpeI-HF (NEB, cat. no. R3133S) was accordingly used. For CRISPR/Cas9 knockout, lentiCas9-Blast (Addgene, plasmid no. 52962) was the source of Cas9. gRNA sequences were designed with the CRISPick tool (https://portals.broadinstitute.org/gppx/crispick/public) and plasmids were produced by mutagenesis of pLenti-gRNA-puro using primers in **Supplementary File 5** as described previously[3]. Plasmids for mammalian expression of all forms of 3XFLAG-tagged TDG (TDG-CD, TDG-FL, TDG-N140A-CD, TDG-N140A-FL) were individually constructed by gene synthesis (Integrated DNA Technologies (IDT)) flanked by a 5' BamHI site and 3' EcoRI site (sequences in **Supplementary File 6**). These restriction sites were then used for cloning in-frame into pLenti- V6.3 Ultra (Addgene, plasmid no. 106172): NEB® Stable Competent *E. coli* (C3040H) were used to produce all lentiviral plasmids. To generate the CMV-eGFP reporter gene, gene synthesis by IDT produced a fragment which encoded (5' to 3') HindIII site, CpG-free eGFP, SV40 polyA, BamHI site, a 160-bp fragment to visualize digestion, NheI site, CpG-free mScarlet, SV40 polyA, and a KpnI site. This fragment was cloned immediately 3' to the CMV promoter in CMV-pCpGl using HindIII and KpnI. Then, a CpG-free CMV-promoter (all CGs mutated to TGs), ordered separately as a gene fragment, was cloned

into this plasmid using BamHI and NheI (gene fragment sequences in **Supplementary File 6**). Sequences were verified by Sanger sequencing.

Assessing sequences approaches for detection of C, 5mC, 5hmC, 5fC, and 5caC

To generate dsDNA containing modified cytosines, a region from the SV40 promoter was amplified using an unmodified forward primer (GCAGGACTAGTGGTGTGGAA) and a reverse primer containing a single 5mC, 5hmC, 5fC, 5caC or unmodified C site (CCATGGACTAAGCTTAGCTCAGAGGC[C]GAGG, where [C] indicates the modified position) purchased from Integrated DNA Technologies, with the following components per 100 µL reaction: 66.75 µL PCR-grade water, 20 µL 5X HF buffer, 2 µL 10 mM dNTPs, 5 µL each 10 µM primer, 0.25 µL 10 ng/µL SV40-pCpGl plasmid, and 1 µL Phusion™ High-Fidelity DNA Polymerase (Thermo Fisher, cat. no. F553L). PCR was performed with a 30 s initial denaturation step at 98 °C followed by 35 cycles of 10 s at 98 °C and 30 s at 72 °C, and a final extension for 5 min at 72 °C. Amplicons were then purified with the Monarch® PCR & DNA Cleanup Kit according to the manufacturer's instructions. Standard bisulfite conversion was performed with the EZ DNA Methylation-Gold Kit (Zymo Research, cat. no. D5005) according to the manufacturer's protocol and using 1 µg DNA as input. Enzymatic methyl conversion was performed with the NEBNext® Enzymatic Methyl-seq Conversion Module (NEB, cat. no E7125L) according to manufacturer's instructions and using 200 ng DNA as input. APOBEC conversion is described below. Oxidative bisulfite conversion was performed as described previously[240]. Briefly, 1 µg DNA was prepared by passing through a Micro Bio-Spin P-6 SSC Columns (Bio-Rad, cat. no. 7326201) that had been washed four times with PCR-grade water. 1 µg DNA was then denatured in a 19 µL reaction supplemented with 0.95 µL 1 M NaOH by incubation at 37 °C for 30 min with shaking and cooling 5 min on ice. 1 µL Potassium perruthenate(VII) (15 mM $KRuO_4$ in 0.05 M NaOH, Alfa Aesar, cat. no. 11877) was then added to the denatured DNA and the reaction was incubated on ice for 1 h and vortexed every 20 min. A second denaturation was performed by the addition of

4 µL of 0.05 M NaOH and incubation again at 37 °C for 30 min with shaking, followed by cooling 5 min on ice. 1 µL of $KRuO_4$ solution (15 mM in 0.05 M NaOH) was added to the reaction and incubated on ice for 1 h and vortexed every 20 min. Oxidized DNA was then purified with the Monarch® PCR & DNA Cleanup Kit and eluted in 20 µL: the entire volume was then bisulfite-converted with the EZ DNA Methylation-Gold Kit (as described above). For 5caC detection by chemical modification-assisted bisulfite conversion (CAB-seq[245]), 1 µg of DNA was mixed with 85 µL Buffer 1 (20 mM N-hydroxysuccinimide, 2 mM 1-ethyl-3-(3-dimethylaminopropyl)carbodiimide hydrochloride, 75 mM 2-(N-morpholino)ethanesulfonic acid (MES buffer) pH = 5) and incubated at 37 °C for 1 h. The DNA was then purified with the Monarch® PCR & DNA Cleanup Kit and eluted in 20 µL Buffer 2 (100 mM sodium phosphate, 150 mM NaCl, 10 mM p-Xylylenediamine) and incubated at 37 °C for 1 h. The DNA was purified again with the Monarch® PCR & DNA Cleanup Kit and eluted in 20 µL: the entire was volume was then bisulfite-converted with the EZ DNA Methylation-Gold Kit. Methylase-assisted bisulfite conversion (MABS) was performed by two rounds of methylation with M.SssI (NEB, cat. no. M0226L) according to the manufacturer's protocol and using 1 µg DNA as input. DNA was cleaned after each round with the Monarch® PCR & DNA Cleanup Kit and eluted in 20 µL PCR-grade water. After the second round of methylation, the entire 20 µL eluant was used for bisulfite conversion with the EZ DNA Methylation-Gold Kit. C/T ratios were then calculated with PCR and pyrosequencing, described below.

Pyrosequencing

PCR for pyrosequencing was performed using HotStarTaq DNA Polymerase (QIAGEN, cat. no. 203205). Each sample was amplified in a 25 µL reaction containing 18.92 µL PCR-grade water, 2.5 µL 10X Buffer, 0.75 µL 25 mM $MgCl_2$, 0.5 µL 10 mM dNTP mix, 0.5 µL each 10 µM primer, 1 µL converted DNA, and 0.33 µL HotStarTaq DNA polymerase. Primers for each target are listed in **Supplementary File 5**. Cycling conditions were 15 min at 95 °C for 15 min for initial denaturation, followed by 55 cycles

of 30 s at 95 °C, 30 s at various annealing temperatures (determined previously with a temperature gradient), and 30 s at 72 °C, followed by 5 min final extension at 72 °C. 20 µL PCR product was prepared for pyrosequencing by the addition of 40 µL PyroMark Binding Buffer (QIAGEN), 19 µL PCR-grade water, and 1 µL Streptavidin Sepharose High Performance beads (Cytiva, cat. no 90100484). Samples were incubated at room temperature at 1400 rpm for at least 10 min. 3 µL 3 µM sequencing primers (listed in **Supplementary File 5**) were diluted in 27 µL PyroMark Annealing Buffer. Pyrosequencing samples were processed in the PyroMark Q24 instrument and vacuum workstation according to the manufacturer's instructions and run-specific protocols designed by the PyroMark Q24 software (QIAGEN).

In vitro methylation and oxidation of plasmid DNA

400 µg each of CMV-pCpGl plasmid and SV40-pCpGl plasmid were methylated by 4 rounds of methylation with M.SssI CpG methyltransferase (NEB, cat. no. M0226S). For the first round, methylation of each plasmid was performed in 4 x 1 mL reactions, each containing 100 µL 10X buffer (NEBuffer 2) 15 µL M.SssI (4,000 units/mL), 25 µL 32 mM S-adenosylmethionine (SAM), 100 µg plasmid, and water to a final volume of 1 mL. The reactions were incubated at 37 °C overnight. The reactions were cleaned by the addition of 20 µL 20 mg/mL Proteinase K, incubation at 55 °C for 2 h, and purified with the Monarch® PCR & DNA Cleanup Kit according to manufacturer's protocol. Total elution volume was 860 µL such that M.SssI reaction components were again added for a second round of methylation. This was repeated for a total of 4 overnight methylation reactions and final elution was performed with 500 µL PCR-grade water. DNA concentrations were measured by NanoDrop™ 2000 (Thermo Fisher, cat. no. ND-2000) and methylation was verified with bisulfite conversion with the EZ DNA Methylation-Gold Kit (as described above) followed by PCR and pyrosequencing with primers listed in **Supplementary File 5**. Methylated plasmid yields were approximately 200 µg (50%). "Mock" methylation controls were used to produce unmethylated plasmid DNA that went through the same steps as methylation except for the addition of M.SssI and SAM.

Methylated DNA was then oxidized with components from the NEBNext® Enzymatic Methyl-seq Conversion Module (NEB, cat. no. E7125L) with a modified procedure. For each 50 µL reaction, 500 ng methylated plasmid DNA was diluted in 29 µL water. Then, on ice, the following components were added to the DNA mix: 10 µL reconstituted TET2 reaction buffer, 1 µL oxidation supplement, 1 µL dithiothreitol (DTT) solution, and 4 µL TET2 protein. Oxidation enhancer was not added to the reaction. The reaction was mixed thoroughly by pipetting followed by the addition of 5 µL diluted Fe(II) Solution, mixing, and incubation overnight at 37 °C, after which 1 µL Stop Reagent was mixed into the reaction and incubated for an additional 30 min at 37 °C. The oxidized DNA was purified with the Monarch® PCR & DNA Cleanup Kit. This reaction was scaled 100X to oxidize 50 µg methylated DNA that was used for all experiments. Oxidation was verified with bisulfite conversion (described above) and APOBEC conversion (described below) followed by PCR and pyrosequencing. 500 ng CMV-pCpGl plasmid per 50 µL TET2 reaction was optimized based on a dose curve which showed negligible changes in CpG oxidation efficiency between 200 and 500 ng with a reduction in non-CpG oxidation efficiency of 8% (**Supplementary Figure 18**).

APOBEC-conversion

DNA was prepared differently for APOBEC conversion depending on DNA type. Unmethylated, methylated and oxidized plasmid DNA was diluted to 200 ng in 16 µL PCR-grade water to be used directly for APOBEC conversion. 16 µL ChIP DNA prepared according to manufacturer's instructions with the ChIP-IT High Sensitivity Kit (Active Motif, cat. no. 53040) was used directly for APOBEC conversion, but input DNA from the corresponding samples were further cleaned as follows: 200 ng DNA (measured with NanoDrop) was diluted in 50 µL PCR-grade water and cleaned with the Monarch® PCR & DNA Cleanup Kit (NEB, cat. no. T1030L) according to manufacturer's protocol. Briefly, 250 µL binding buffer was mixed with diluted DNA and loaded onto a column. The column was centrifuged for 30 s at 16,000 x *g* and the sample was washed

twice by the addition of 200 µL DNA wash buffer followed by centrifugation as before. The second centrifugation was extended to 60 s and DNA was eluted into a clean 1.5 mL tube in 17 µL PCR-grade water. For genomic DNA extracted by the phenol:chloroform method, 200 ng DNA was diluted in 100 µL PCR-grade water, sonicated on a Bioruptor (Diagenode) for 3 min at high power (30 s on/30 s off) and cleaned as described above with the Monarch® PCR & DNA Cleanup Kit, adjusting the binding buffer volume to 500 µL. APOBEC conversion was then performed with the NEBNext® Enzymatic Methyl-seq Conversion Module. Briefly, 4 µL formamide was added to 16 µL of prepared DNA on ice in 200 µL PCR tubes and incubated for 10 minutes at 85 °C in a preheated thermocycler with the heated lid set to 90 °C. The denatured DNA was placed immediately on ice, cooled for 2 min, and mixed with 80 µL prepared master mix containing 68 µL PCR-grade water, 10 µL APOBEC reaction buffer, 1 µL BSA, and 1 µL APOBEC enzyme per sample. The reaction was mixed thoroughly and incubated at 37 °C for 3 h in a thermocycler with the heated lid set to 45 °C. The reaction was then purified as described previously with the Monarch® PCR & DNA Cleanup Kit and eluted in 10 µL PCR-grade water.

Transfection of luciferase reporter plasmids

For luciferase assays, 120,000 cells were plated per well of a 6-well tissue-culture dish 24 h prior to transfection. For ChIP and/or CpG oxidation/methylation detection by pyrosequencing, 1.2 million HEK293 cells were plated in a 100-mm dish 24 h prior to transfection. Transfection was performed with X-tremeGENE™ 9 DNA Transfection Reagent (Millipore Sigma, cat. no 6365787001). Per each well of a 6-well plate, 50 ng luciferase reporter plasmid was diluted in 50 µL Opti-MEM™ I Reduced Serum Medium (Thermo Fisher, cat. no. 31985062) in a 1.5 mL tube. In a separate 1.5 mL tube, 1 µL X-tremeGENE 9 was mixed into 50 µL Opti-MEM. The Opti-MEM/DNA mixture was added into the X-tremeGENE 9/Opti-MEM mixture and incubated for 20 min at room

temperature. All ~100 µL was then added in a dropwise manner to each well of cells. Reactions were scaled to create master mixes for all transfected wells. For 100-mm dishes, all volumes were increased by 10-fold. Cells were incubated with transfected DNA until the time points indicated for each experiment.

Luciferase assays

Luciferase assays were performed using the Luciferase Assay System (Promega, cat. no E4550). At indicated time points after transfection, the media was removed and 200 µL 1X Reporter Lysis Buffer was added to each well. Membranes were disrupted with one freeze-thaw cycle by briefly transferring cell culture dishes to -80 °C, followed by collection with a cell scraper and clarification by centrifugation at 13,000 x *g* for 1 min. 10 µL (CMV-pCpGl) or 50 µL (SV40-pCpGl) clarified lysate was diluted to a final volume of 100 µL in 1X Reporter Lysis Buffer in 5-mL round-bottom polystyrene tubes. Immediately prior to quantification of each sample, 50 µL re-suspended Luciferase Assay Reagent was added to each sample and light emission was measured immediately in a Monolight 3010 luminometer (Analytical Luminescence Laboratory). Sample protein concentrations were determined by the Bradford Protein Assay (Bio-Rad, cat. no. 5000006) and A595 readings were measured in a DU 730 UV–Vis Spectrophotometer (Beckman Coulter), using a bovine serum albumin standard curve, and luciferase activity was normalized to concentration.

Lentivirus production

24 h prior to transfection, HEK293T cells were plated at a density of 3.8 x 10^6 per 100 mm dish. Cells were transfected using X-tremeGENE™ 9 DNA Transfection Reagent as described above. Briefly, individual lentiviral transfer plasmids were mixed with a packaging plasmid (pMDLg/pRRE, Addgene #12251), envelope protein plasmid (pMD2.G, Addgene #12259), REV-expressing plasmid (pRSV-Rev, Addgene #12253),

and a 3:1 ratio (transfection reagent : total DNA) of X-tremeGENE™ 9 DNA Transfection Reagent in Opti-MEM medium (Gibco). The mixture was incubated for 20 min at room temperature and added in a drop-wise manner to HEK293T cells in 8 mL of fresh DMEM medium in a 100 mm dish. Lentiviral particles were harvested by filtering the supernatant through a 0.45 μm disk filter 72 h after transfection and either used immediately or stored at −80 °C. 10 μg/mL Blasticidin S HCl was used to select for stable transformants.

Mouse tissue and human cells

Surgically excised cerebral cortices of adult (14 weeks) C57BL/6J female mice were purchased from the Jackson Laboratory. HEK293, HEK293T, HepG2, MCF-7, MDA-MB-231, T24, RKO, SKHep1, and HLEC cell lines were purchased from ATCC. Normal human fibroblasts (IMR90) were purchased from the Coriell Institute (cat. no. I90-15). Normal hepatocyte and breast cell lines were purchased from Celprogen (cat. nos. 33003-02, 77002-07). All cell lines were maintained in DMEM, high glucose (Thermo Fisher, cat. no. 11965118) supplemented with 10% fetal bovine serum (FBS) (Wisent, cat. no. 080-150) and 1X Penicillin-Streptomycin-Glutamine (Thermo Fisher, cat. no. 10378016).

Gene knockout and TDG N140A mutation with CRISPR/Cas9

For knockout experiments, 10,000 HEK293 cells were seeded in a 100-mm tissue-culture dish in 10 mL of complete DMEM and transfected 24 h later with Cas9 and gRNA expression plasmids. Transfection was performed using X-tremeGENE™ 9 DNA Transfection Reagent by mixing 250 ng Cas9 plasmid and 250 ng gRNA plasmid in 100 μL Opti-MEM™ I Reduced Serum Medium. To produce the N140A substitution in TDG, 200 pmol of an 89-bp single-stranded donor oligo (ssODN) with ~40-bp homology arms was also mixed with Cas9/gRNA plasmids (**Supplementary File 5**). Separately, 1.5 μL

X-tremeGENE™ 9 was diluted into 200 µL Opti-MEM™. Both solutions were mixed thoroughly by pipetting and the DNA solution was added in a dropwise manner to the solution containing X-tremeGENE™ 9. The mixture was incubated at room temperature for 20 minutes and then added in a dropwise manner to the 100-mm dish containing the HEK293 cells. 48 h after transfection, cells were trypsinized and re-plated at a density of 200 cells per 100-mm dish in 4 dishes. The cells were allowed to expand (~30 days), replacing the media only after 20 days, in order to form sufficiently large colonies. 20-30 individual colonies were then picked manually with a sterile P200 micropipette tip and moved to individual wells of 24-well plates. The clones were left to expand for ~10 days, trypsinized, and each split to three wells of a 6-well plate. When cells reached 80-90% confluence, one well was collected for freezing, one for DNA isolation, and one for protein isolation. For freezing, cells were washed once with 500 µL phosphate-buffered saline (PBS) and detached in 400 µL trypsin for 5 minutes at 37 °C. Trypsinization was neutralized by the addition of 1 mL complete DMEM followed by collection of the cells into 1.5-mL tubes, centrifugation at 300 x *g* for 5 min at 4 °C, aspiration of the supernatant, resuspension of the cells in ice-cold freezing solution (80% complete DMEM, 10% FBS, 10% DMSO), and storage at -80 °C. For DNA isolation, cells were washed with 500 µL PBS prior to the addition of 500 µL DNA lysis buffer (100 mM Tris, pH 7.5, 150 mM NaCl, 0.5% SDS, 10 mM EDTA) directly to the well. 400 µL of lysed cells were transferred to a 1.5 mL tube and processed by RNAse A treatment, proteinase K treatment, phenol-chloroform extraction, and ethanol precipitation as described previously[2]. 100 ng of purified DNA was used as a template to amplify the regions surrounding the gRNA binding sites with Q5® High-Fidelity 2X Master Mix (NEB, cat. no. M0492L) according to the manufacturer's protocol. Sanger sequencing was performed with forward or reverse PCR primers at the Genome Québec Centre d'expertise et de services. For *TDG* mutations, PCR amplicons were further subcloned into pCR®4-TOPO TA vector as described above and Sanger sequenced with M13R primer to determine individual allelic DNA sequences. For all other knockouts, sequencing of mixed (i.e., no subcloning) PCR products served as an initial screen to

prioritize clones for protein quantification by western blot. Exceptionally for *MBD2* and *MBD3* knockouts, gRNA sites were in GC-rich regions that failed to amplify despite several different approaches; therefore, these knockout cell lines were screened directly by western blot. For protein isolation, DMEM was aspirated from the wells and cells were detached directly in 1 mL of PBS by pipetting. Cell suspensions were transferred to 1.5 mL tubes and centrifuged at 300 x *g* for 5 min at 4 °C followed by aspiration of the supernatant and re-suspension in 50 µL RIPA buffer (20 mM Tris, 2 mM EDTA, 150 mM NaCl, 1% Triton X, 0.1% SDS, 0.5% deoxycholate). Lysed cells were incubated for 30 minutes on ice and lysates were then clarified by centrifugation at 13,000 x *g* for 10 min at 4 °C. The supernatants were transferred to new 1.5 mL tubes and stored at -80 °C.

Western blotting

Protein concentrations were quantified with the Bradford Protein Assay using a bovine serum albumin (BSA) standard curve and A595 readings were measured in a DU 730 UV–Vis Spectrophotometer (Beckman Coulter). 30 µg protein was mixed with 2X Laemmli Sample Buffer (Bio-Rad, cat. no. 1610737), boiled for 5 min, and separated on a 12% acrylamide gel (with a 5% acrylamide stacking layer). Gels were run for 10 min at 110 V and then for 50 min at 170 V, transferred onto nitrocellulose membranes for 90 min at 250 mA, blocked for 30 min with 5% skim milk powder in TBST (20 mM Tris, 150 mM NaCl, 0.1% Tween® 20, pH=7.4), and incubated with primary antibody in 5% skim milk powder in TBST overnight at 4 °C. Membranes were rinsed 5 times with TBST, incubated with secondary antibody diluted in 5% skim milk powder in TBST, and rinsed 5 more times with TBST. All antibodies and dilutions are listed in **Supplementary Table 2**. After addition of Clarity Western ECL Substrate (Bio-Rad, cat. no. 1705061), images were acquired with semiautomatic exposure on the Amersham Imager 600.

Flow cytometry

Cells were collected by trypsinization and re-suspended in PBS containing 1% FBS for sorting at 5 x 10^6 cells per mL. Cell sorting was performed on a Becton Dickinson FACSAria Fusion equipped with 4 lasers (405nm; 488nm; 561nm and 635nm). Debris were excluded by analyzing a fluorescence-negative control HEK293 cell lines in Forward Scatter (FSC) against Side-Scatter (SSC). Doublets were excluded by analyzing FSC-Area versus FSC-Height. Single color controls were used to calculate the spectral spillover of the fluorescent proteins. Cells were counted/sorted based on their single or double positive expression of fluorescent proteins mScarlet and eGFP. The sort precision mode was set to purity and the sort efficiency was maintained above 95%. A 100uM nozzle was used and a pressure of 20 psi was applied to the system. Cells were sorted into 15-mL conical tubes containing complete DMEM. For APOBEC-seq analysis, cells were pelleted by centrifugation at 300 x g for 3 min at 4 °C, re-suspended in 400 µL DNA lysis buffer and processed as described above for DNA isolation, APOBEC conversion, and pyrosequencing.

Chromatin immunoprecipitation

Chromatin immunoprecipitation of mouse cortices and human cell lines was performed with the ChIP-IT High Sensitivity® Kit (Active Motif, cat. no 53040) according to the manufacturer's protocol. Frozen mouse cortices were cut into ~1 mm cubes in 10 mL Complete Tissue Fixation Solution in a 100-mm petri dish and incubated, with rotation, at room temperature for 15. The reaction was quenched with 515 µL Stop Solution followed by incubation at room temperature for 5 minutes. The cross-linked tissues were then homogenized by passing the suspension 20 times through a 20-guage needle attached to a sterile plastic syringe. The tissues were pelleted by centrifugation at 1,250 x *g* for 3 min at 4 °C and pellets were washed twice with 10 mL ice-cold PBS Wash Buffer. Washed pellets were resuspended in 5 mL Chromatin Prep Buffer supplemented with 5 µL protease inhibitor cocktail (PIC) and 5 µL 100 µM PMSF, incubated on ice for 10 minutes, and homogenized again to lyse cell membranes. Nuclei were pelleted by

centrifugation as before, resuspended in 1 mL ChIP buffer, incubated for 10 min on ice, and sonicated in a Bioruptor (Diagenode) sonicator in 200 µL aliquots, each receiving 6 10-min rounds of sonication at high power 4 °C at 16,000 x g and stored at -80 °C. Input DNA was prepared as follows: 25 µL was removed and purified according to the manufacturer's protocol (briefly, RNAse A treatment, Proteinase K treatment, addition of NaCl and reverse cross-linking at 65 °C for 16 hours, and purification by phenol:chloroform and ethanol extraction) and quantified using a NanoDrop spectrophotometer. 30 µg sheared chromatin was immunoprecipitated overnight at 4 °C by dilution in 200 µL ChIP Buffer supplemented with 5 µL PIC and incubation with 4 µg antibody with 5 µL Blocker (see **Supplementary Table 2**). 30 µL washed Protein G agarose beads were added to the reaction and incubated for 3 h. The reactions were then diluted by the addition of 600 µL ChIP buffer and added to ChIP filtration columns, washed 5 times with Wash Buffer AM1, and eluted in 100 µL Elution Buffer AM4. ChIP DNA was purified in the same manner as input DNA and in 36 µL elution volume for ChIP-sequencing or 200 µL for ChIP-qPCR. Cultured cells in 150-mm tissue-culture dishes were processed in a similar manner and as described in the manufacturer's protocol for cultured cells.

qPCR and RT-qPCR

For RT-qPCR, RNA was isolated from 80-90% confluent 100-mm tissue-culture dishes by resuspension of washed cells in 1 mL of Trizol reagent (Thermo Fisher, cat. no. 15596018) and RNA extraction was performed according to the Trizol manufacturer protocol. Briefly, 200 mL of chloroform was added to 1 mL of Trizol/RNA mixture. The samples were thoroughly vortexed, incubated at room temperature for 2 min, and centrifuged for 15 min at 12,000 x *g* at 4 °C. The aqueous phase was transferred to a new 1.5 mL tube prior to the addition of 0.5 mL isopropanol and incubation at room temperature for 10 min. The samples were centrifuged for 10 min at 12,000 x *g* at 4 °C and washed twice with 75% ethanol, discarding the supernatant each time. The pellets were air dried for 10 min and resuspended in 50 µL PCR-grade water. Concentrations

were measured with the RNA BR Assay (Thermo Fisher, cat. no. Q10211) and 1 µg RNA was used for each reverse transcriptase reaction using M-MuLV Reverse Transcriptase (NEB) according to manufacturer protocol. cDNA was diluted 1:2 (20 µL reverse transcription reaction to 40 µL water) for RT-qPCR. ChIP DNA was used directly as eluted from the the ChIP-IT High Sensitivity® Kit for qPCR. 2 µL of diluted cDNA or ChIP DNA was amplified in the LightCycler ® 480 Instrument II (Roche) in a 20 µL reaction containing 10 µL LightCycler ® 480 SYBR Green I Master Mix (Roche, cat. no. 04887352001) and 0.8 µL each of 10 µM forward and reverse primer listed in **Supplementary File 5**. Quantification was performed by Roche Lightcycler Software.

Treatment with pharmacological inhibitors

For treatment with pharmacological inhibitors, 1.2 x 10^6 HEK293 cells were plated on 100-mm tissue-culture dishes. Cells were transfected with 500 ng oxidized CMV-pCpGl plasmid 24 hours after plating as described above. 3 hours after transfection, the media were replaced with pre-mixed media containing the pharmacological inhibitors at the indicated concentrations. Actinomycin D (Millipore Sigma, cat. no. A9415), 5,6-Dichlorobenzimidazole 1-β-D-ribofuranoside (DRB) (Millipore Sigma, cat. no. D1916), and cycloheximide (Millipore Sigma, cat. no. C4859) were initially dissolved in sterile DMSO while amanitin (Millipore Sigma, cat. no. A2263) was dissolved in sterile PCR-grade water. The cells were collected 24 h after transfection by cross-linking and were then processed for ChIP as described above. For western blot, cells were plated and treated in the same way but with no transfection and collected in 100 µL RIPA buffer and processed for western blot as described above. For the trypan blue exclusion test, cells were treated as for western blot, but resuspended at 1:1 ratio with 0.4% Trypan Blue Solution (Thermo Fisher, cat. no 15250061) (100 µL each). Cells were counted under a light microscope using a hematocytometer; viable cell count was determined as the number of total cells minus blue staining cells and dead cell count was determined as the number of blue staining cells.

Native FLAG-TDG-N140A-CD protein co-immunoprecipitation and LC-MS/MS

Native co-immunoprecipitation of FLAG-TDG-N140A-CD was performed according to a previously published protocol for nuclear proteins[523]. 1.5 x 10^7 HEK293, HepG2, or MCF-7 cells stably expressing 3XFLAG-TDG-N140A-CD or control (pLenti- V6.3 Ultra empty vector) were plated on 150-mm tissue-culture dishes in 20 mL of complete DMEM. 24 h later, the cells were collected with a cell scraper, transferred to 50-mL conical tubes, centrifuged at 300 x *g* for 3 min at 4 °C, washed once with 10 mL ice-cold PBS, centrifuged, resuspended in 1 mL of ice-cold PBS, transferred to 1.5 mL tubes, centrifuged, resuspended in 950 µL RSB (10 mM Tris-HCl, pH 7.4, 5 mM MgCl2, 10 mM NaCl) supplemented with 1X final concentration cOmplete™, Mini, EDTA-free Protease Inhibitor Cocktail (Millipore Sigma, cat. no. 11836170001) and incubated on ice for 15 min. 20 µl of 10% Triton™ X-100 (Millipore Sigma, cat. no. X100) was then added and the reactions were centrifuged at 1,200 x *g* for 5 min at 4 °C. The supernatant containing the cytoplasmic fraction was discarded and the nuclear pellet was resuspended in 200 µL Lysis Buffer 1[523] and processed in a Bioruptor sonicator for 10 min at low power (30 sec on/60 sec off) and then incubated on an end-to-end rotator for 1 h at 4°C. The tubes were then centrifuged at 8,000 x *g* at 4 °C for 10 min and the supernatant was collected into a separate 1.5 mL tube. The pellet was then processed in this way with 50 µL Lysis Buffers 2 and 3 and 150 µL Lysis Buffer 4, each time incubating for 30 min on an end-to-end rotator at 4°C, centrifuging, and collecting the supernatant. At the point of collection 50 µL, 50 µL, and 100 µL RSB were added to the supernatant containing Lysis Buffers 2, 3, and 4, respectively. The four extracts were pooled and centrifuged for clarification at 4 °C for 5 min at 3,000 x *g*. 50 µL pooled extract was retained to be used as input and the remaining pooled extract was used for immunoprecipitation. 100 µL Anti-FLAG® M2 Magnetic Beads (Millipore Sigma, cat. no M8823-5ML) per sample were washed twice with 500 µL RSB, the pooled extract was added to the washed beads, and the reactions were incubated on an end-to-end rotator

at 4 °C for 2 h. The beads were then washed 5 times with Wash Buffer (10 mM HEPES-NaOH, pH 7.4, 10% glycerol, 165 mM NaCl, protease inhibitor). For western blotting, proteins were eluted from the beads by the addition of 100 µL 2X Laemmli Sample Buffer and boiling for 10 min. Supernatants were collected with a magnetic rack and diluted 1:1 in RSB for western blotting. For LC-MS/MS, washed beads (prior to the elution step) were provided directly to the Research Institute of the McGill University Health Center (RI-MUHC) Proteomics Platform. For each sample, protein complexes on beads were loaded onto a single stacking gel band to remove lipids, detergents and salts. The gel band was reduced with DTT, alkylated with iodoacetic acid and digested with trypsin. Extracted peptides were re-solubilized in 0.1% aqueous formic acid and loaded onto a Thermo Acclaim Pepmap (Thermo Fisher, 75 µM ID X 2cm C18 3 µM beads) precolumn and then onto an Acclaim Pepmap Easyspray (Thermo Fisher, 75 µM X 15cm with 2 µM C18 beads) analytical column for separation using a Dionex Ultimate 3000 uHPLC at 250 nl/min with a gradient of 2-35% organic (0.1% formic acid in acetonitrile) over 2 hours. Peptides were analyzed using a Thermo Orbitrap Fusion mass spectrometer operating at 120,000 resolution (FWHM in MS1) with HCD sequencing (15,000 resolution) at top speed for all peptides with a charge of 2+ or greater. The raw data were converted into .mgf files (Mascot Generic Format) for searching using the Mascot 2.6.2 search engine (Matrix Science) against human protein sequences (Uniprot 2022). The database search results were loaded onto Scaffold Q+ Scaffold_4.9.0 (Proteome Sciences) for statistical treatment (independent-t-test) and data visualization. Proteins significantly enriched in FLAG-TDG samples were used as input for visualization and testing for significantly enriched gene ontology categories with STRING (https://string-db.org/).

ChIP-seq library preparation and sequencing

All library preparation and sequencing was performed by the NGS Services team of Genome Quebec. Libraries were generated from 25 µL of fragmented DNA (range 100-

300 bp) using the NEBNext Ultra II DNA Library Prep Kit for Illumina (NEB), as per the manufacturer's recommendations. Adapters and PCR primers were purchased from Integrated DNA Technologies (IDT). Size selection was carried out using SparQ beads (QIAGEN) prior to PCR amplification (12 cycles). Libraries were quantified using the KAPA Library Quantification Kits - Complete kit (Universal) (Kapa Biosystems). Average fragment size was determined using a LabChip GX (PerkinElmer) instrument. The libraries were normalized, pooled, and then denatured in 0.05 N NaOH and neutralized using HT1 buffer. The pool was loaded at 175 pM (HEK293 cells) or 200pM (mouse cortices) on a Illumina NovaSeq S4 lane using Xp protocol as per the manufacturer's recommendations. The run was performed for 2x100 cycles (paired-end mode). A phiX library was used as a control and mixed with libraries at 1% level. Base calling was performed with RTA v3.4.4 . Program bcl2fastq2 v2.20 was then used to demultiplex samples and generate fastq reads.

ChIP-seq data processing

Paired-end FASTQ files were trimmed for quality and adapter content using Trim Galore v0.6.7 with default parameters and then aligned to human (hg38) or mouse (mm39) reference genomes using bowtie2[456] v2.4.5 with default parameters. Resulting SAM files were converted to BAM format and unmapped reads were removed with samtools[524] v1.16.1. BAM files were then sorted and duplicates were marked with Picard tools v2.18.29 and indexed with samtools. To generate bigWig files (used for plotting ChIP-seq signal over genes and regions) BAM files were converted to bedGraph using the bamCoverage function of deepTools[525] v3.5.1 with the argument --normalizeUsing RPKM. These files were then sorted with bedtools[526] v2.30.0 and converted to bigWig files using the bedGraphtoBigWig tool[527], which were used to generate ChIP-seq signal plots using the deepTools computeMatrix and plotHeatmap functions. Significant peaks were identified with the callpeak function from macs2[457] v2.2.7.1 using the respective input sequencing data for each sample as control and

precompiled values for the mappable human and mouse genome sizes. karyoploteR [528]was used to plot ChIP-seq signal and peaks in **Figure 5**. Observed/expected ratios and p-values for TDG and MBD3 binding peaks in gene regions in mouse cortices were calculated directly with the annotatePeaks.pl function in homer v4.11[529]. For mouse SNPs and INDELs, all bam files from all ChIP and input samples were merged, SNPs and INDELs were called with freebayes v1.3.6, and then filtered for QUAL scores above 30 using vcffilter v1.0.3. The custom HEK293 reference genome (used for APOBEC-seq read alignment) was built by merging all ChIP-seq alignments (BAM files) from anti-FLAG-TDG/input samples with samtools, calling SNPs and INDELs with freebayes v1.3.6, filtering for SNPs with QUAL scores above 30 using vcffilter v1.0.3, and applying called SNPs (no INDELs to avoid changes to chromosomal coordinates) to the hg38 genome using the Genome Analysis Toolkit (GATK) v4.2.6.1 FastaAlternateReferenceMaker tool.

APOBEC-seq library preparation and sequencing

Libraries were generated from 200 ng fragmented input or 3.5-200 ng ChIP DNA (prepared for APOBEC conversion as described above) using the NEBNext® Enzymatic Methyl-seq Kit (NEB, cat. no. E7120L) with some modifications to the manufacturer's protocol. End prep and ligation of EM-seq adapter ligation were performed as per the manufacturer's protocol except that the NEBNext EM-seq Adapter was substituted with 5hmC Adapter provided as a gift by NEB: this is critical as methylated NEBNext Em-seq Adapters included with the kit are not resistant to APOBEC-conversion alone. Cleanup after adapter ligation was performed with NEBNext Sample Purification Beads as per manufacturer's protocol but elution volume was reduced to 17 µL for compatibility with the modified protocol. The TET2 oxidation step of the NEBNext® Enzymatic Methyl-seq Kit protocol was then skipped and 16 µL eluant was used directly for the APOBEC conversion (denaturation and deamination) steps. Denaturation with formamide, deamination with APOBEC, cleanup, PCR amplification, and final cleanup steps were

performed according to the manufacturer's protocol. Libraries were quantified using the KAPA Library Quanitification Kits - Complete kit (Universal) (Kapa Biosystems). Average size fragment was determined using a LabChip GX (PerkinElmer) instrument. The libraries were normalized and pooled and then denatured in 0.05N NaOH and neutralized using HT1 buffer. The pool was loaded at 200pM on a Illumina NovaSeq S4 lane using Xp protocol as per the manufacturer's recommendations. The run was performed for 2x150 cycles (HEK293 cells) or 2x100 cycles (mouse cortices) (paired-end mode). A phiX library was used as a control and mixed with libraries at 1% level. Base calling was performed with RTA v3.4.4. The program bcl2fastq2 v2.20 was then used to demultiplex samples and generate fastq reads.

APOBEC-seq data processing

Paired-end FASTQ files were trimmed for quality and adapter content using Trim Galore v0.6.7 with default parameters and then aligned to custom HEK293 (hg38) or mouse (mm39) reference genomes using bismark v0.23.1 with default parameters. Bismark alignments were processed by Bismark deduplication and methylation extractor scripts. Conversion rates were calculated with BCREval[31]. BedGraph files were filtered for indicated coverage thresholds and converted to bigwig files for visualization with deepTools as described above for ChIPseq data. Genomic region annotations of CpGs were performed with the annotatePeaks.pl function in homer (v4.11) or overlapped with significant peaks of histone marks with the bedtools intersect function. To calculate enrichment/depletion of oxidized CpGs in specific regions, observed oxidized CpG counts in each region were compared to expected counts calculated from all CpGs that passed coverage thresholds regardless of oxidation status by the Fisher's exact test using the phyper function in R and p values were adjusted for multiple testing with the Bonferroni correction. Gene ontology enrichment analyses were performed and visualized with ShinyGO v0.77[530]. Histograms in **Supplementary Figure 12** were generated with the methylKit[531] package in R.

Infinium MethylationEPIC Array

To ensure sufficient DNA quantity for the Infinium MethylationEPIC Array, 200 ng genomic DNA was prepared for and converted by APOBEC as described above in three technical replicates per biological replicate for a total of 600 ng converted DNA. DNA was measured and concentrated by SpeedVac to a 20 µL final volume. The DNA was processed for the array according to the manufacturer's instructions. Briefly, 6 µL was used for amplification, followed by overnight incubation at 37 °C, fragmentation for 1 h at 37 °C, precipitation and resuspension in RA1, and denaturation (95 °C, 20 min then cooling at room temperature for 30 min) prior to hybridization (30 µL) with the array overnight at 48 °C. BeadChips were washed, stained and scanned with the Illumina iScan System and processed with Illumina GenomeStudio (2011) software. Ultimately, no significantly differentially methylated probes were identified between HEK293 and HEK293 TDGKO cells using multiple independent t-tests with correction by the FDR method. No highly oxidized sites were discovered after applying filters for HEK293-specific SNPs and INDELs generated in this study.

14.6 Data Availability

ChIP-seq data generated in this study have been deposited in the Gene Expression Omnibus (GEO) under the accession numbers: GSE228704 (human) and GSE228706 (mouse). APOBEC-seq data generated in this study have been deposited under the accession numbers GSE228703 (human) and GSE228705 (mouse). Raw mass spectrometry data have been deposited to the UCSD MassIVE database under the dataset identifier MSV000091458. Public data used in this experiment are available from ENCODE under the accession numbers: ENCFF301UTR (HEK293 H3K4me1 ChIP-seq), ENCFF451UZW (HEK293 H3K27ac ChIP-seq), ENCFF046YRR (HEK293 H3K4me3 ChIP-seq), ENCFF496OIF (HEK293 H3K36me3 ChIP-seq), ENCFF037SXA (HEK293 H3K9me3 ChIP-seq), ENCFF235UTX (MCF-7 pol2 ChIP-seq),

ENCFF978ENR (MCF-7 MTA1 ChIP-seq), ENCFF187EDY (MCF-7 HDAC2 ChIP-seq), ENCFF313LLN (MCF-7 GATAD2B ChIP-seq), ENCFF448ZOJ (IMR90 pol2 ChIP-seq), ENCFF354VWZ (HepG2 pol2 ChIP-seq) ENCFF858LTF/ENCFF626IQB (mouse cortex WGBS). HEK293 ENCODE pol2 ChIP-seq data was obtained from GEO under the accession GSE31477. HEK293 small RNA-seq data was obtained from GEO under the accession GSE137834. HEK293 RNA-seq data was obtained from GEO under the accession GSE139420. HEK293 MBD3 ChIP-seq data was obtained from GEO under the accession GSE102945 and TET1 and TET2 ChIP-seq data was obtained from GEO under the accession GSE172141.

14.7 Acknowledgements

We acknowledge the RIMUHC Proteomics and Molecular Analysis Platform for assistance with protein analysis and the Centre d'expertise et de services Génome Québec and the Montreal Clinical Research Institute (IRCM) for next-generation sequencing services. We also thank Julien Leconte and Camille Stegen for performing flow cytometry at the McGill University Flow Cytometry Core Facility and acknowledge the Digital Research Alliance of Canada and Calcul Québec for computing and storage resources. We also thank Farida Vaisheva and Sergiy Dymov for assistance in protein isolation from knockout cell lines. Finally, we thank New England Biolabs for their gift of 5hmC adapters. This study was funded by the Canadian Institutes of Health Research (PJT159583). D.M.S was supported by fellowships from the McGill University Faculty of Medicine and Health Sciences (Friends of McGill Fellowship; JP Collip Fellowship in Medical Research; James Frosst Fellowship).

14.8 Author Contributions

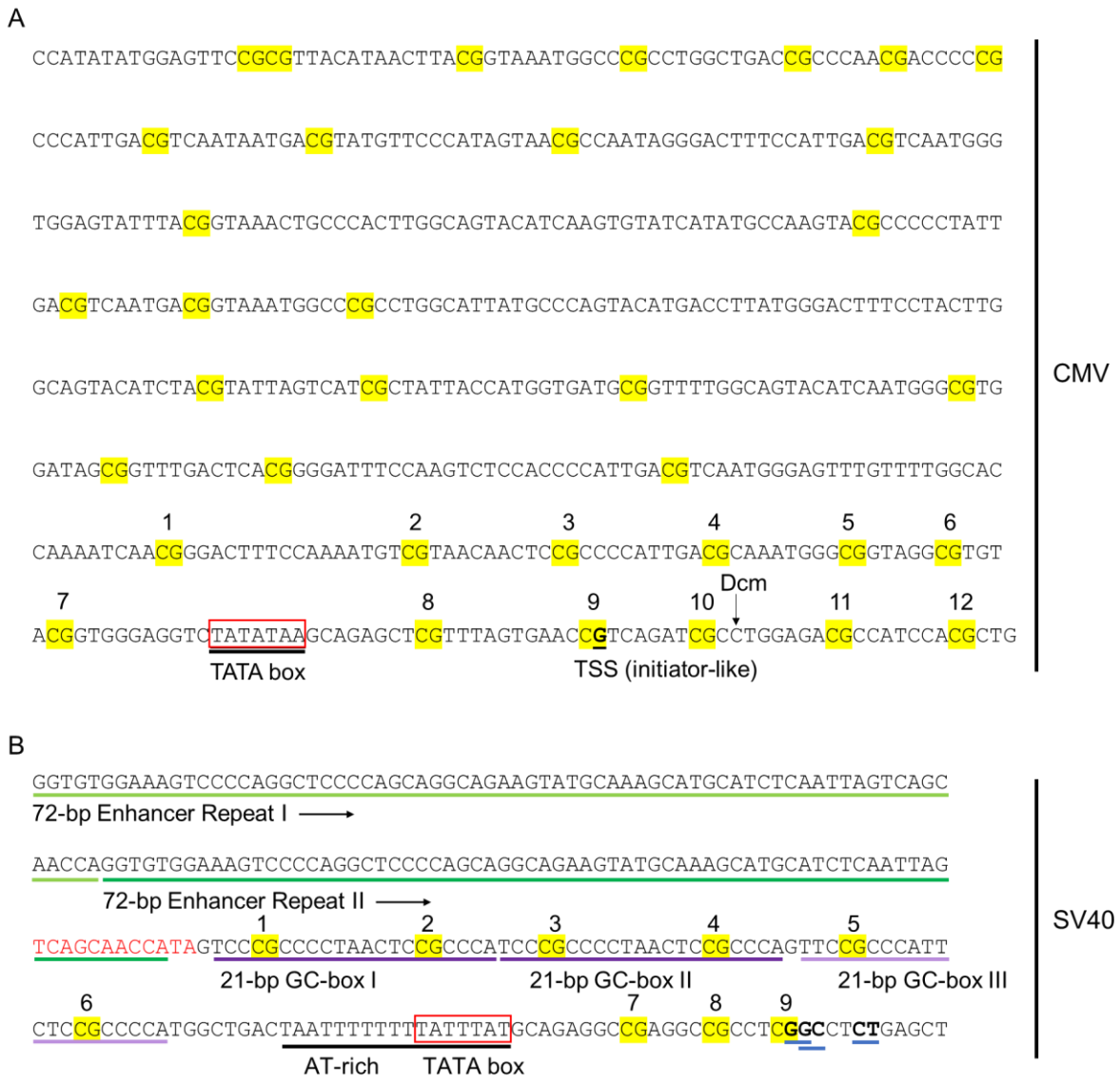
D.M.S. and M.S. designed the experiments, planned analyses, and wrote the manuscript. D.M.S. performed all experiments and bioinformatics analyses.

14.9 Competing Interests

The authors have no competing interests to declare.

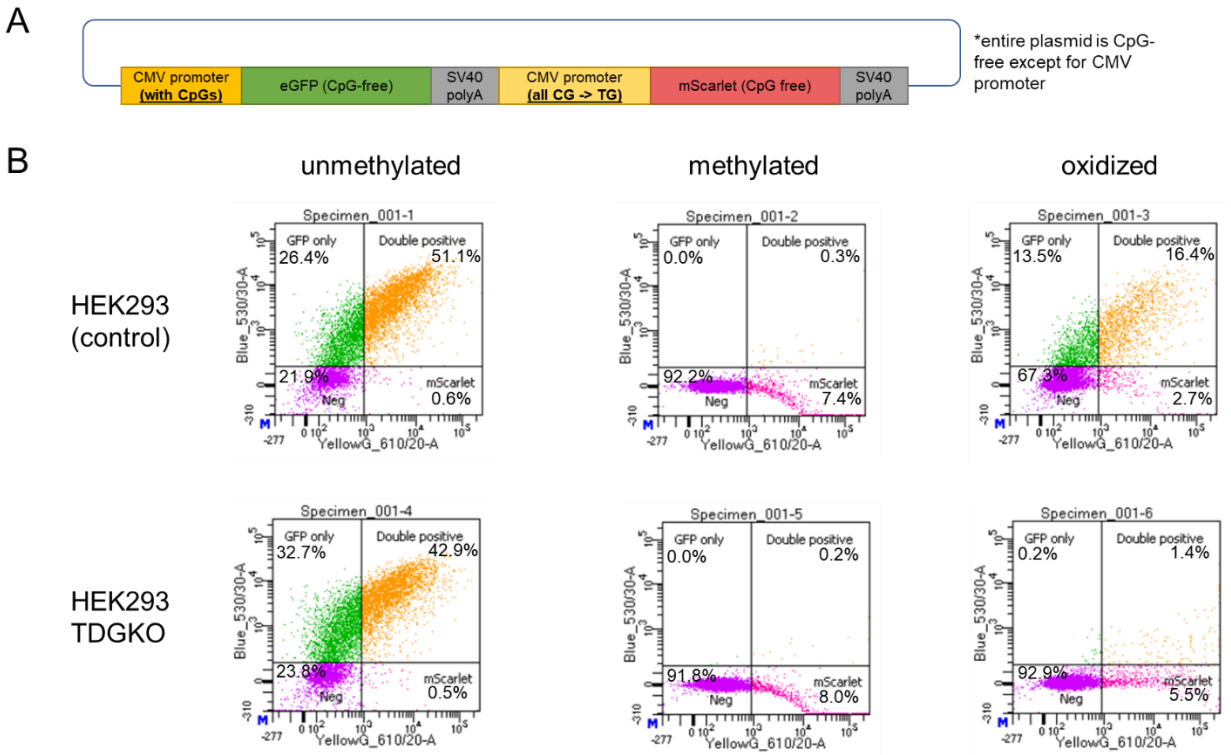
14.10 Supplementary Information

This supplementary information contains Supplementary Figures 1-18 and Supplementary Tables 1-2. Supplementary Files 1-6 can be accessed via the following link: https://drive.google.com/drive/folders/1Uy7M-HCgtCG_N308xe3CrYFNP4-beJ9G?usp=sharing

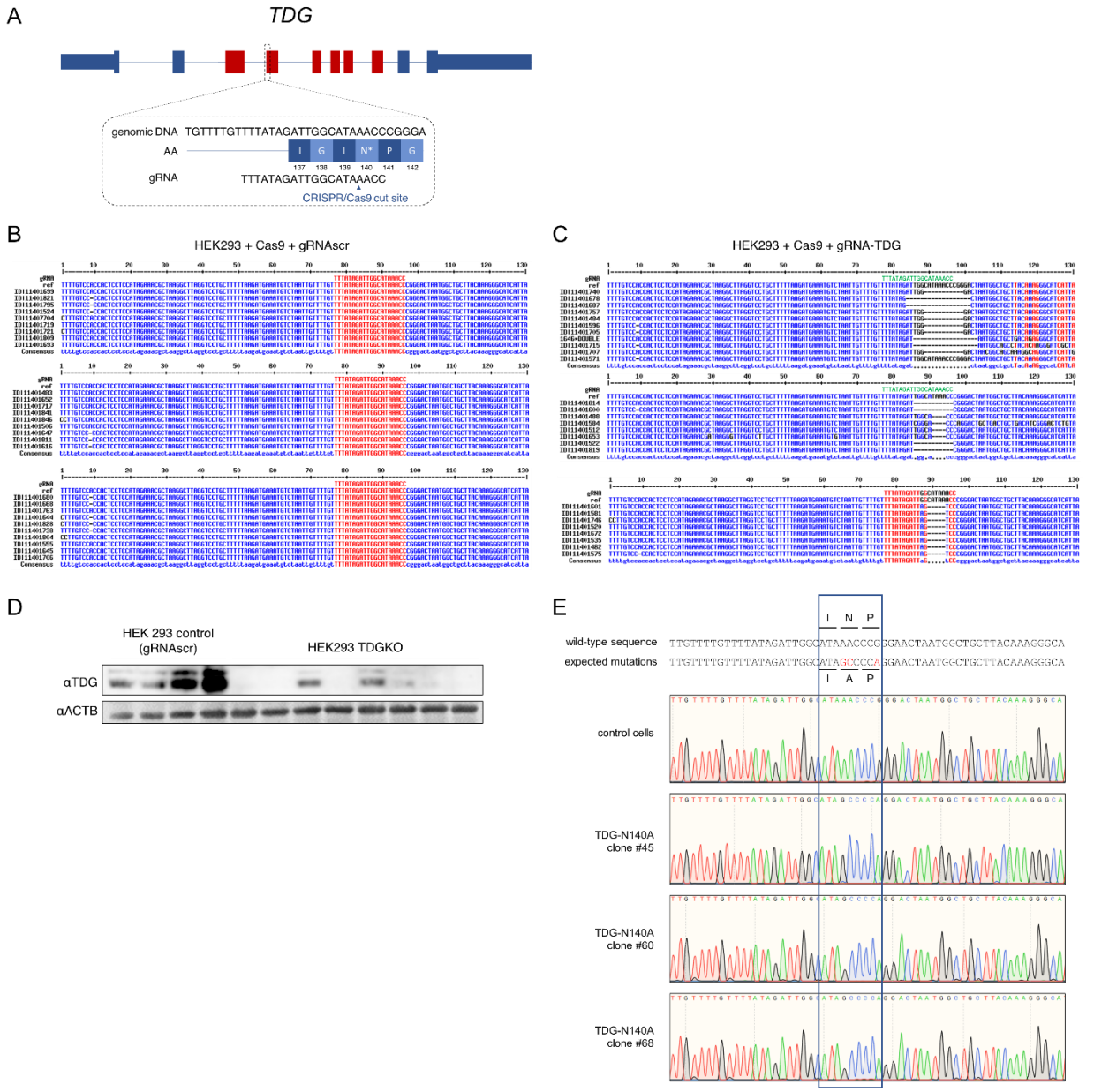


Supplementary Figure 1. Sequences of promoter-reporter plasmids used in this study. (A) The sequence of the CMV promoter that was cloned into the pCpGl plasmid. CpGs are highlighted in yellow, the TATA box and TSS are marked and CpGs near the TSS are numbered as they are referenced in the study. (B) The sequence of the SV40 promoter and CpG-less enhancer repeats that were cloned into the pCpGl plasmid. Elements relevant to transcriptional activity are marked, including: the TATA box, an A-T

rich region near the TATA box, 3 GC-boxes, two 72-bp enhancer repeats, and three major TSS. All CpGs are numbered as they are referenced in the study.

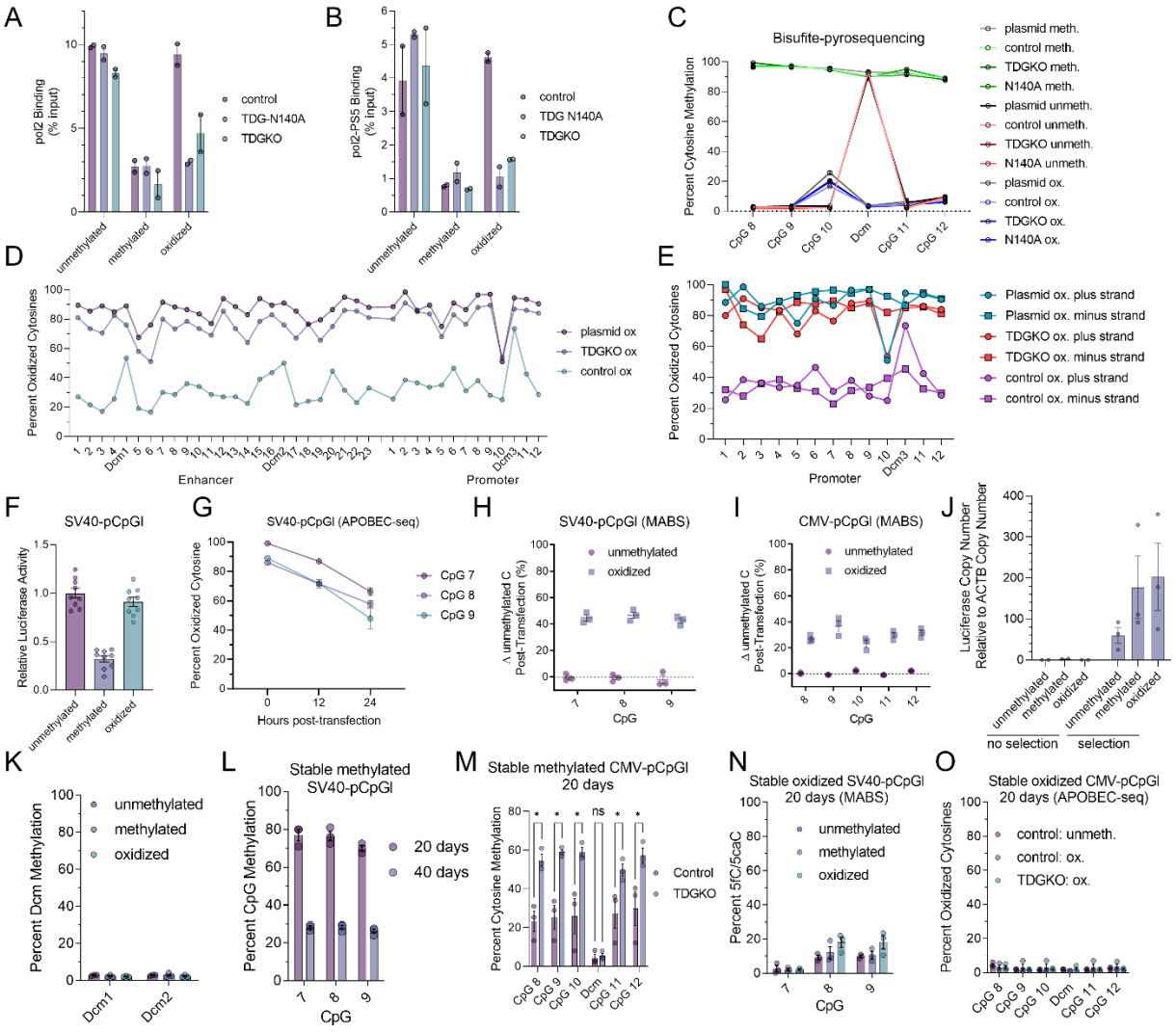


Supplementary Figure 2. Replication of luciferase results by flow cytometry. (A) Schematic diagram of the plasmid used for flow cytometry experiments. It contains a CMV promoter as in CMV-pCpGL which can be methylated or oxidized, followed by a CpG-free eGFP reporter. A CpG-free CMV promoter also drives expression of CpG-free mScarlet to assess transfection efficiency, though detection is hampered by low expression rate of the CpG-free CMV promoter. SV40 polyA terminators follow both genes. (B) Flow cytometry scatter plots showing mScarlet signal (x-axis) and eGFP signal (y-axis) in HEK293 control cells and HEK293 TDGKO cells transfected with 500 ng reporter plasmid from (A) and subjected to flow cytometry 24 h post-transfection.



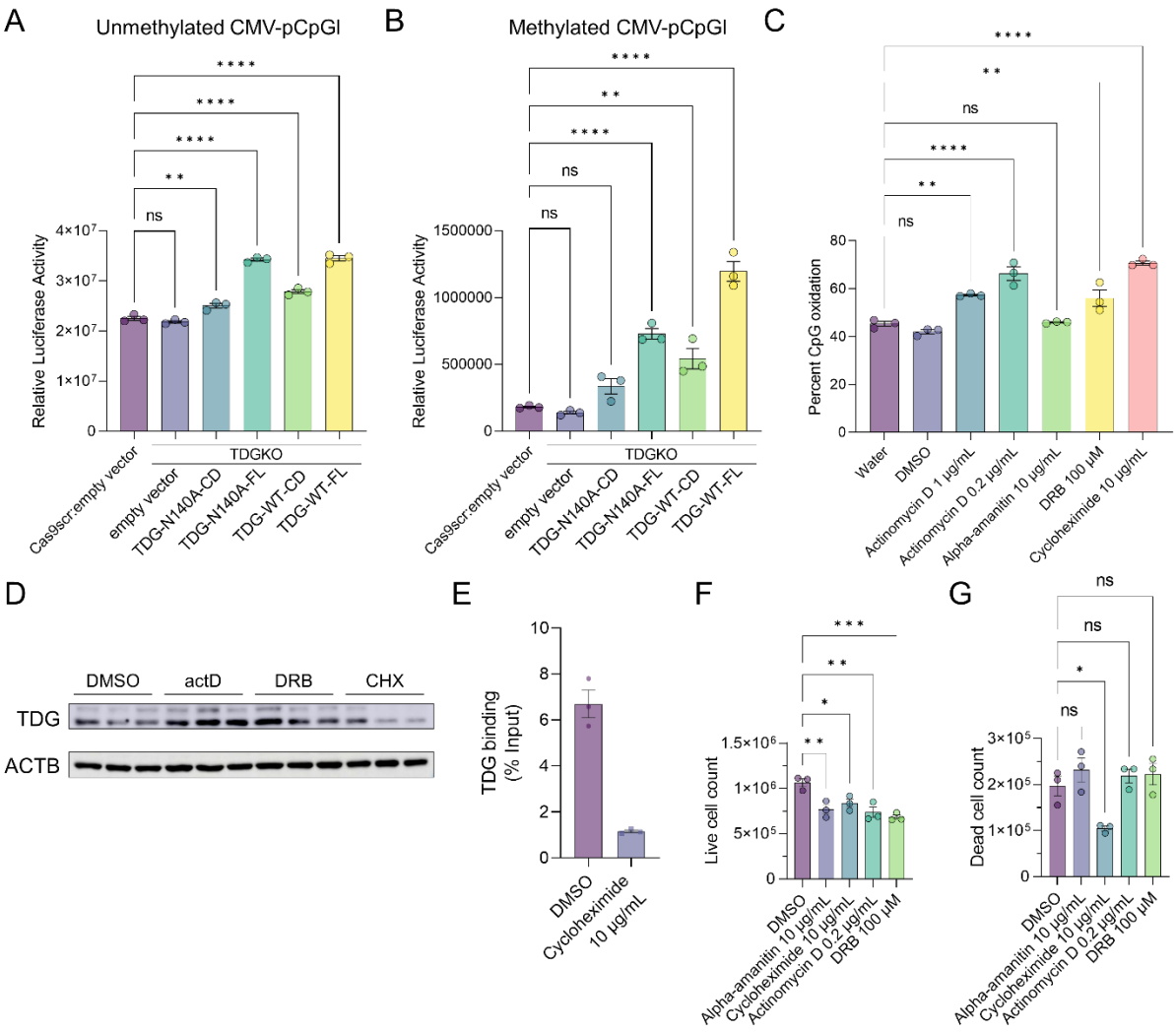
Supplementary Figure 3. TDG knockout and mutagenesis by CRISPR/Cas9. (A) Schematic diagram of the gene structure of *TDG* (not to scale). The box highlights the region that is targeted by the CRISPR/Cas9 guide RNA, which is predicted to cleave the DNA immediately adjacent to the catalytic N140 residue, indicated with an asterisk. (B-C) Sanger sequencing results aligned to the reference genome and gRNA for 3 HEK293 clonal cell lines per condition that were transfected with Cas9 and (B) a scrambled gRNA (gRNAscr) or (C) a gRNA targeting TDG: 8-11 subcloned PCR fragments are

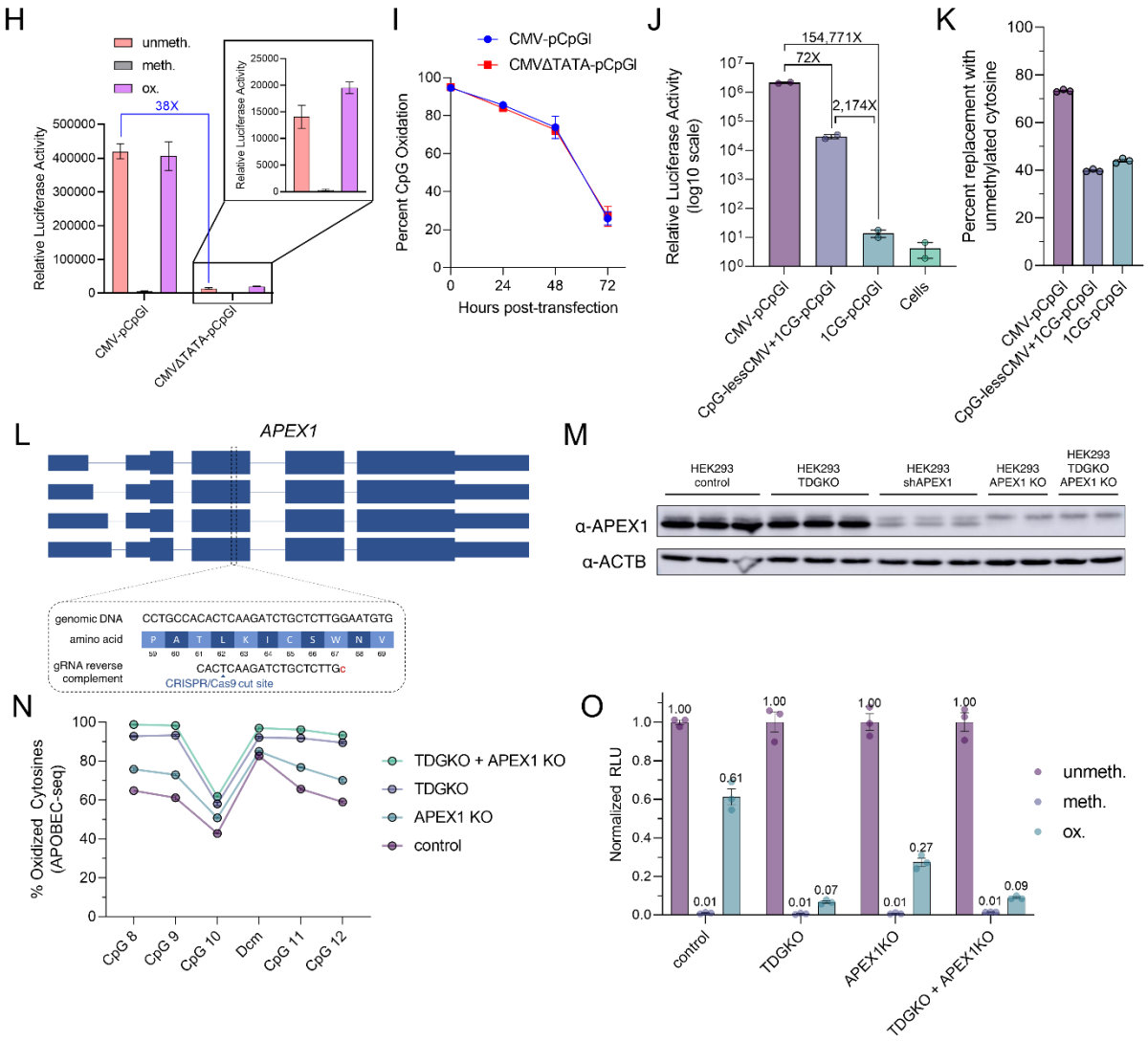
shown per clonal cell line. In (C) different mutation profiles all produce frame-shift mutations and occasionally result in misalignment of the short gRNA sequence, which is replaced in green when this occurs. (D) Western blot using antibody against TDG and ACTB (loading control) depicting 4 control HEK293 clonal cell lines transfected with Cas9/gRNAscr and 8 HEK293 clonal cell lines transfected with Cas9/gRNA-TDG, prioritized for screening by western blot based on Sanger sequencing results. Clones with no TDG signal were selected as TDG knockout cell lines for downstream experiments. (E) Sanger sequencing chromatograms of one control cell line (transfected with Cas9/gRNAscr) and three successfully mutated clonal cell lines transfected with Cas9/gRNA-TDG and ssODN (see Methods) which show the intended mutation profile and no additional mutations. The reference human genome (wild-type) sequence for TDG is depicted and includes the neighboring (5') intronic sequence as shown in (A), with the first three amino acids of the exon highlighted by a blue box. In the sequence below, the mutations expected to be produced by CRISPR/Cas9 are highlighted in red and the corresponding original and expected amino acid sequences are also shown. The additional silent mutation in proline removes the CRISPR/Cas9 protospacer adjacent motif (PAM) so that once the N140A mutation is introduced, the site can no longer be targeted by CRISPR/Cas9.



Supplementary Figure 4. Additional analyses of oxidized cytosine dynamics in HEK293 cells. (A) RNA polymerase II binding (expressed as percent input and measured by ChIP-qPCR) of each form of CMV-pCpGl plasmid 48 h post-transfection into control, TDGKO, and TDG-N140A HEK293 cells. (B) Same as (A) but using an antibody against the form of RNA polymerase II phosphorylated on serine 5. (C) Percent methylated cytosines determined by bisulfite-pyrosequencing of each condition in as in Figure 2G across 5 CpGs and one non-CpG cytosine (Dcm). Plasmid labels indicate original methylation levels in the three plasmids that were used for transfection. (D) A complete profile of all 35 CpGs and 3 non-CpG cytosines (Dcm1-3) in the oxidized

CMV-pCpGl plasmid 48 h post-transfection, measured by APOBEC-pyrosequencing of pol2-bound DNA from (A). The promoter CpGs 8-12 and Dcm3 correspond to those measured by APOBEC-seq in the majority of the experiments in this study. (E) As in (D), APOBEC-pyrosequencing results of pol2-bound DNA with assays that sequence both plus and minus strands of the DNA. (F) Relative luciferase expression of unmethylated, methylated, or oxidized 50 ng SV40-pCpGl plasmid 48 h post-transfection into HEK293 cells, normalized to the unmethylated condition. (G) APOBEC-pyrosequencing results of oxidized SV40-pCpGl plasmid as in (F) as a function of indicated time points post-transfection. (H) MAB-(pyro)sequencing results of unmethylated or oxidized SV40-pCpGl 48 h post-transfection into HEK293 cells. Data are presented as the change in unmethylated cytosine percent, due to overestimation of oxidized cytosines inherent to MABS. (I) Same as (H) but CMV-pCpGl was transfected in place of SV40-pCpGl. (J) Detection of luciferase DNA by qPCR of each form of transfected CMV-pCpGl plasmid with or without 20 days of selection for stable integration by antibiotic resistance. (K) DNA methylation assessed by bisulfite-pyrosequencing of two Dcm sites in methylated CMV-pCpGl after 20 days of selection. (L) DNA methylation assessed by bisulfite-pyrosequencing of 3 CpGs in methylated SV40-pCpGl after 20 or 40 days of selection for stable integration with antibiotic. (M) DNA methylation assessed by bisulfite-pyrosequencing of 5 CpGs and Dcm site in methylated CMV-pCpGl after 20 days of selection for stable integration in control HEK293 cells or TDGKO cells. (N) Percent CpG oxidation of 3 CpGs in three forms of stably integrated (20 days selection) SV40-pCpGL assessed by MABS. (O) Percent CpG oxidation assessed by APOBEC-seq of 5 CpGs and Dcm site in CMV-pCpGl after 20 days of selection for stably integrated plasmid in control or TDGKO cells. Data are presented as mean ± SEM, with individual replicates plotted as circles in all bar graphs; in line plots, circles represent the mean, in which n=3 biological replicates. Multiple t-tests with correction for multiple testing were conducted in GraphPad Prism v9.4.1. * indicates $p<0.05$ and ns indicates no statistically significant difference.

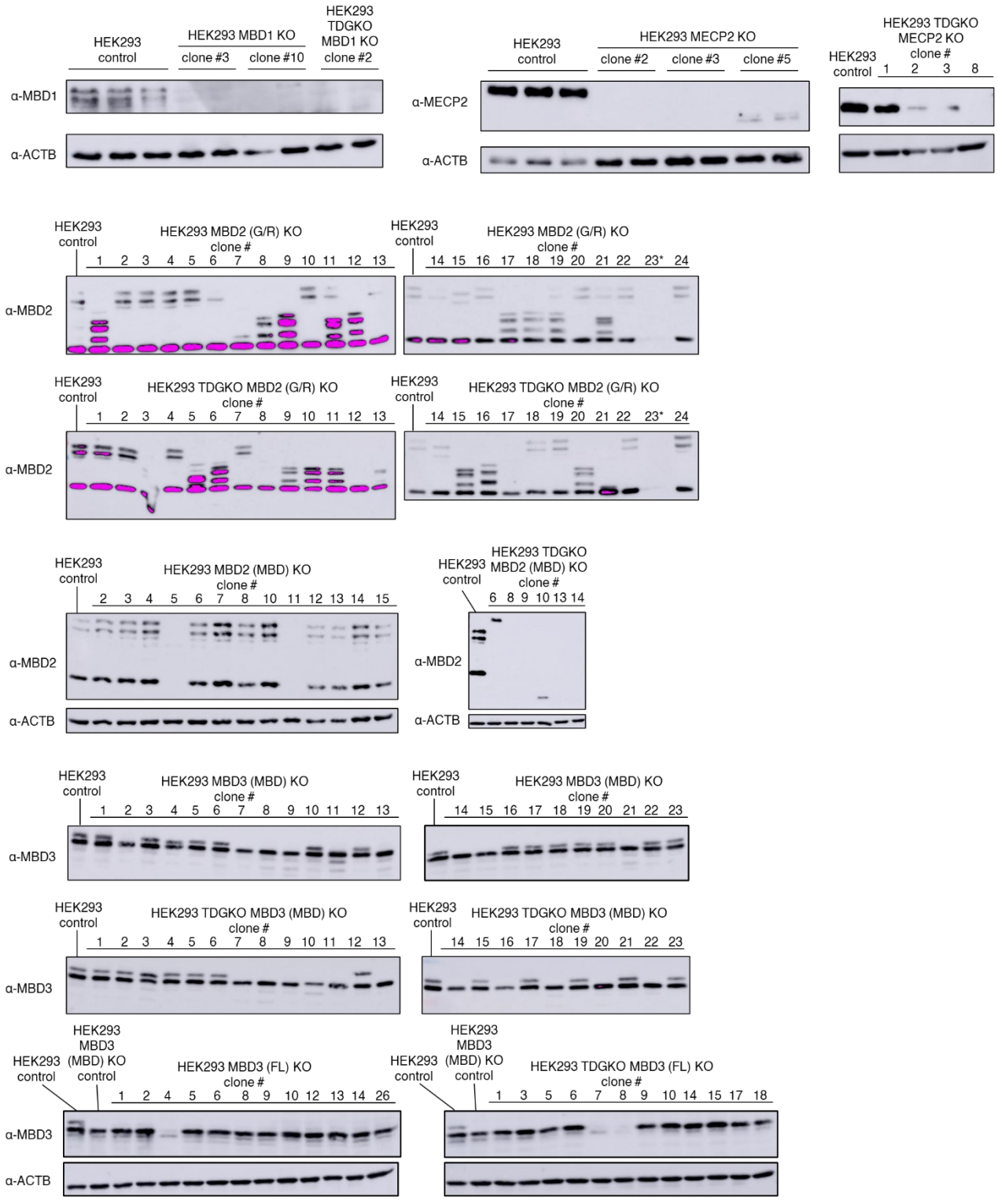

A
Unmethylated CMV-pCpGl
Relative Luciferase Activity
4×10^7
3×10^7
2×10^7
1×10^7
0
****
****
****
**
ns
Cas9scr:empty vector
TDGKO
empty vector
TDG-N140A-CD
TDG-N140A-FL
TDG-WT-CD
TDG-WT-FL
B
Methylated CMV-pCpGl
Relative Luciferase Activity
1500000
1000000
500000
0
****
**
****
ns
ns
Cas9scr:empty vector
TDGKO
empty vector
TDG-N140A-CD
TDG-N140A-FL
TDG-WT-CD
TDG-WT-FL
C
Percent CpG oxidation
80
60
40
20
0
****
**
ns
****
**
ns
Water
DMSO
Actinomycin D 1 µg/mL
Actinomycin D 0.2 µg/mL
Alpha-amanitin 10 µg/mL
DRB 100 µM
Cycloheximide 10 µg/mL
D
DMSO
actD
DRB
CHX
TDG
ACTB
E
TDG binding (% Input)
10
8
6
4
2
0
DMSO
Cycloheximide 10 µg/mL
F
Live cell count
1.5×10^6
1×10^6
5×10^5
0
***
**
*
**
DMSO
Alpha-amanitin 10 µg/mL
Cycloheximide 10 µg/mL
Actinomycin D 0.2 µg/mL
DRB 100 µM
G
Dead cell count
3×10^5
2×10^5
1×10^5
0
ns
ns
*
ns
DMSO
Alpha-amanitin 10 µg/mL
Cycloheximide 10 µg/mL
Actinomycin D 0.2 µg/mL
DRB 100 µM



Supplementary Figure 5. The relationship between transcription and demethylation of oxidized CMV-pCpGI. (A-B) Relative luciferase activity normalized to total protein content in indicated cell lines transfected with 50 ng unmethylated (A) or methylated (B) CMV-pCpGI, corresponding to Figure 3C. (C) Percent CpG oxidation averaged for all 5 CpGs in the CMV-pCpGI APOBEC-pyrosequencing assay after transfection of 50 ng oxidized CMV-pCpGI plasmid into HEK293 cells treated with water or DMSO control or various drugs at indicated concentrations. (D) Western blot of TDG and beta-actin protein levels in response to treatment with DMSO control or pharmacological agents in (C); actD represents actinomycin-D and CHX represents

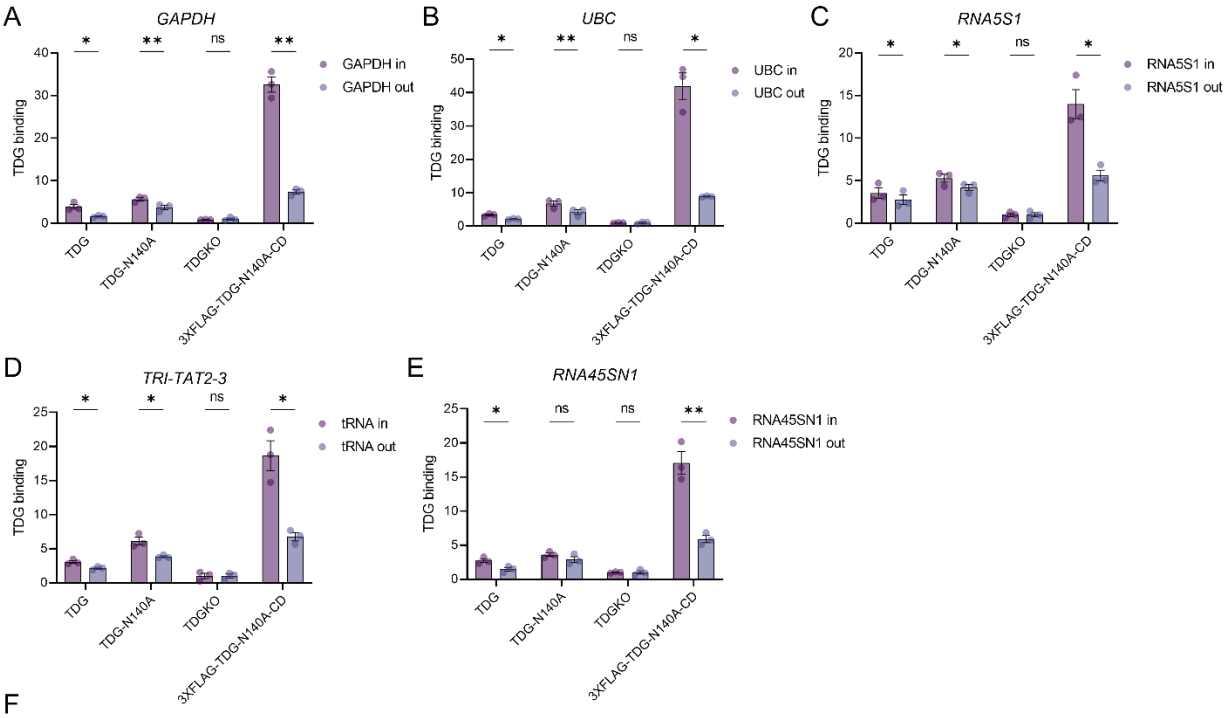
cycloheximide. (E) Endogenous TDG binding to oxidized CMV-pCpGl measured by ChIP-qPCR in cells treated with DMSO (control) or cycloheximide. (F-G) Live and dead cell counts in cells treated with pharmacological inhibitors or DMSO control as measured by the trypan blue exclusion test. (H) Relative luciferase activity normalized to protein content per µL lysate in unmethylated, methylated, or oxidized CMV-pCpGl plasmids compared to similarly modified CMV-pCpGl plasmids with a TATA box mutation. Inset expands on ΔTATA plasmids for better resolution. (I) Averaged percent oxidation in the CMV-pCpGl APOBEC-pyrosequencing assay; CpGs 1-4 in the CMV-pCpGl were averaged since the TATA mutation disrupts the CpG8-12 assay in ΔTATA plasmids. Cells were transfected with oxidized plasmids and collected at the indicated time points after transfection. Time=0 indicates untransfected original plasmid. (J) Relative luciferase activity of unmethylated CMV-pCpGl plasmid and its derivatives 48 h after transfection into HEK293 cells. Axis is in log scale to allow visualization of non-expressing 1CG-pCpGl and untransfected background signal (cells), but linear fold-changes compared to original CMV-pCpGl plasmid are depicted. (K) Percent increase in unmethylated cytosines as detected by APOBEC-pyrosequencing for CpG 12 in CMV-pCpGl or the corresponding single CpG in each of the derived plasmids in experiments in (J). (L) Schematic diagram of APEX1 isoforms, not to scale. The highlighted box focuses on the region that is targeted by CRISPR/Cas9 and shows the DNA sequence, amino acid sequence and positions in the protein, as well as the reverse complement of the APEX1-targeting guide RNA and its corresponding predicted cut site. The lower case and red “c” in the reverse complementary sequence of the guide RNA indicates a 5’ G required for guide RNA expression from the U6 promoter that does not match the genomic target sequence. (M) Western blot results of HEK293 control cells and TDGKO cells and APEX1 knockout clones on each genetic background which were prioritized after Sanger sequencing results. HEK293 lines stably expressing commercially obtained shRNA targeting APEX1 were included for comparison, but were not used in any other experiments in this study due to luciferase signal interference issues apparently stemming from the requirement for stable high shRNA expression. (N) APOBEC-

pyrosequencing results of oxidized CMV-pCpGl plasmid transfected into HEK293 cell lines in (M) 24h post-transfection. (O) Relative luciferase activity normalized to protein content and to each unmethylated condition for each cell line from (M-N), transfected with unmethylated, methylated, or oxidized CMV-pCpGl plasmid and collected 24 h post-transfection. Data are presented as mean ± SEM. Multiple t-tests with correction for multiple testing were conducted in GraphPad Prism v9.4.1. * indicates $p<0.05$, ** indicates $p<0.01$, *** indicates $p<0.001$, **** indicates $p<0.0001$, after multiple correction, and ns indicates no statistically significant difference.

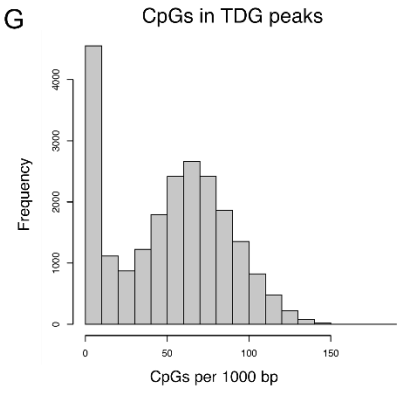


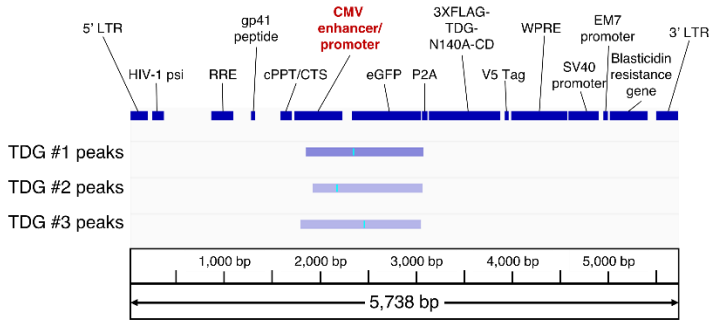
Supplementary Figure 6. CRISPR/Cas9-mediated knockout of MBD proteins in HEK293 and HEK293 TDGKO cell lines. Beta-actin was used as a loading control for

all experiments, except in MBD2 (G/R) and MBD3 (MBD) conditions, where other isoforms serve as a loading control. MBD2 (G/R) refers to cells transfected with the most 5' MBD2 gRNA in Figure 4C and MBD2 (MBD) refers to validated MBD2 (G/R) KO cells that were then transfected with the second gRNA indicated in Figure 4C. Purple coloring in MBD2 (G/R) conditions indicates over-exposure of the western blot which was necessary to visualize fainter MBD2 isoforms. MBD3 (MBD) refers to cells that were transfected with the most 5' MBD3 gRNA in Figure 4C and MBD3-FL (full-length) indicates validated MBD3 (MBD) KO cells that were then transfected with the remaining MBD3 gRNAs. As detailed in the Methods, MBD2 and MBD3 were screened for mutagenesis directly by western blot to the presence of GC-rich sequences that prevented PCR amplification and Sanger sequencing, while other cell lines were initially screened by Sanger sequencing and only prioritized clones were then assessed by western blot.


A
GAPDH
TDG binding
GAPDH in
GAPDH out
B
UBC
UBC in
UBC out
C
RNA5S1
RNA5S1 in
RNA5S1 out
D
TRI-TAT2-3
tRNA in
tRNA out
E
RNA45SN1
RNA45SN1 in
RNA45SN1 out
TDG
TDG-N140A
TDGKO
3XFLAG-TDG-N140A-CD
ns

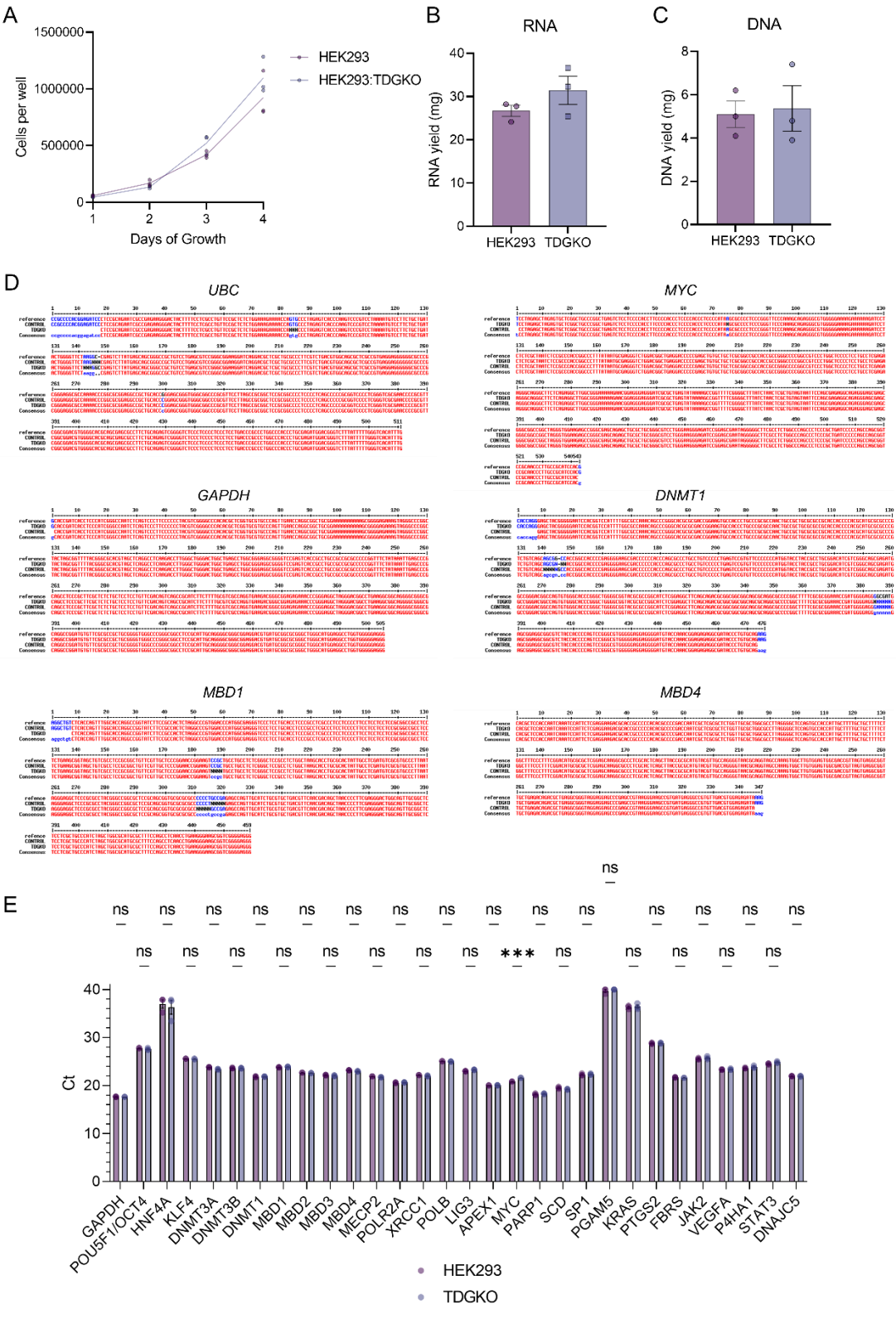

F
Human (hg38)
chr10
pol2 binding
RNA-seq rep. 1
RNA-seq rep. 2
small RNA-seq rep. 1
small RNA-seq rep. 2
small RNA-seq rep. 3
Genes
SEC31B
NDUFB8
HIF1AN
PAX2
pol2 peaks
TDG #1 peaks
TDG #2 peaks
TDG #3 peaks
H3K27ac peaks
H3K4me3 peaks


G
CpGs in TDG peaks
Frequency
CpGs per 1000 bp

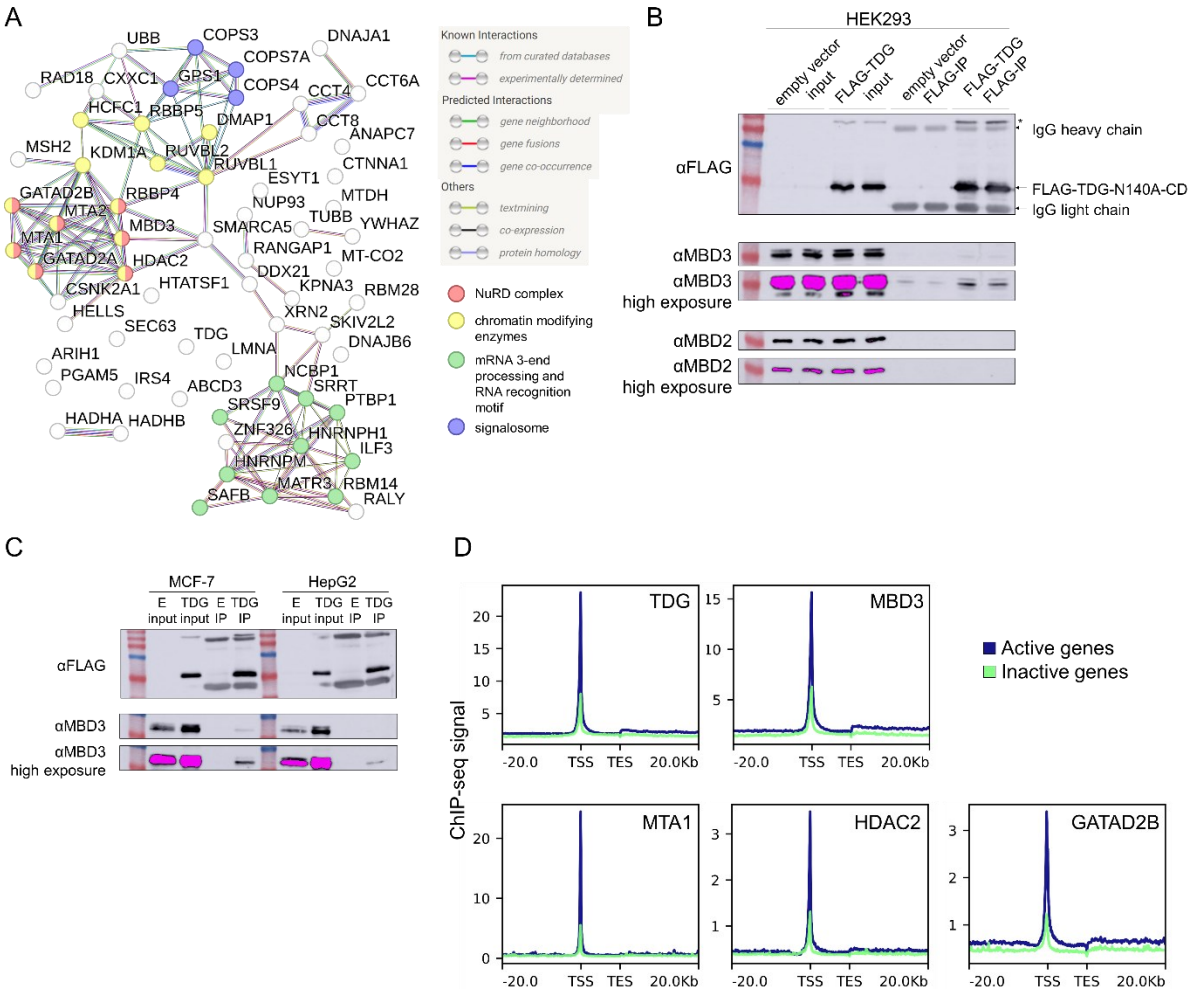

H
5' LTR
HIV-1 psi
RRE
gp41 peptide
cPPT/CTS
CMV enhancer/ promoter
eGFP
P2A
3XFLAG-TDG-N140A-CD
V5 Tag
WPRE
SV40 promoter
EM7 promoter
Blasticidin resistance gene
3' LTR
TDG #1 peaks
TDG #2 peaks
TDG #3 peaks
1,000 bp
2,000 bp
3,000 bp
4,000 bp
5,000 bp
5,738 bp

Supplementary Figure 7. Additional dynamics of TDG binding. (A-E) Validation by ChIP-qPCR of 3XFLAG-TDG-N140A-CD binding to five selected promoters of *GAPDH* (A), *UBC* (B), *RNA5S1* (C), a tRNA gene (*TRI-TAT2-3*, D), and an rDNA repeat (*RNA45SN1*, E). Included also are ChIP-qPCR experiments using antibody against endogenous TDG in HEK293 control cells (TDG), HEK293 TDG-N140A cells (TDG-N140A) and HEK293 TDGKO cells (TDGKO) in the same regions. Primers labeled “in” are those that amplify a region within significant peaks of 3XFLAG-TDG-N140A-CD from ChIP-seq data and primers labeled “out” amplify an adjacent region with no significant 3XFLAG-TDG-N140A-CD peak. Data are normalized to the negative control (TDGKO, “out”) in for each promoter. Individual values are plotted as circles for n = 3 biological replicates, bars indicate mean, and error bars show SEM. FDR-adjusted paired t-tests were used to compare “in” and “out” qPCR values within each replicate after normalization to input, in GraphPad Prism v9.4.1. * indicates $p<0.05$, ** indicates $p<0.01$, after multiple correction, and ns indicates no statistically significant difference. (F) IGV genome browser screenshot of an apparent intergenic enhancer sequence which shows significant 3XFLAG-TDG-N140A-CD peaks in all three replicates of HEK293 ChIP-seq data, as well as significant H3K27ac and H3K4me3 peaks from ENCODE data and absence of RNA-seq signal (public data) but detectable small RNA-seq reads from three replicates (public data) and detectable pol2 signal (ENCODE) which does not constitute a statistically significant peak. (G) Histogram of CpG density across all significant replicated 3XFLAG-TDG-N140A-CD peaks in HEK293 cells. (H) Schematic of the stably integrated lentiviral 3XFLAG-TDG-N140A-CD expression vector, highlighting all elements and demonstrating significant 3XFLAG-TDG-N140A-CD peaks which were identified in the same way as genomic ChIP-seq peaks but using an initial modified hg38 reference genome wherein the depicted sequence was added for alignment. Light blue lines within TDG peaks represent peak summits and are approximately centered at the CMV promoter TSS.

Supplementary Figure 8. Specific tissue-specific 3XFLAG-TDG-N140A-CD peaks in promoters. IGV genome browser screenshots of tissue-specific genes and corresponding significant 3XFLAG-TDG-N140A-CD peaks in MCF-7, IMR90, or HepG2 cells. Light blue lines within TDG peaks represent peak summits.

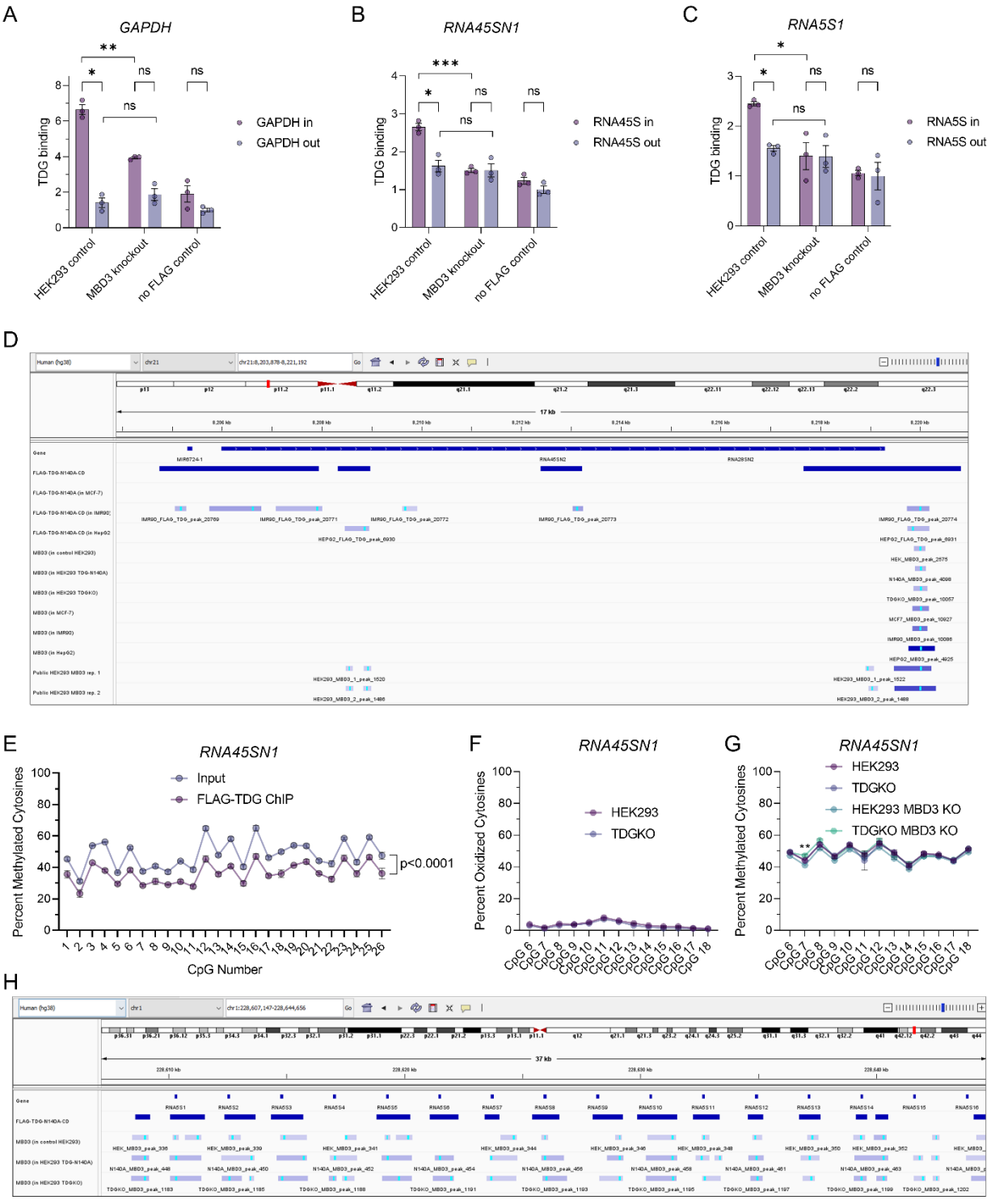

A
Cells per well
1500000
1000000
500000
0
1
2
3
4
Days of Growth
HEK293
HEK293:TDGKO
B
RNA
RNA yield (mg)
40
30
20
10
0
HEK293
TDGKO
C
DNA
DNA yield (mg)
8
6
4
2
0
HEK293
TDGKO
D
UBC
MYC
GAPDH
DNMT1
MBD1
MBD4
E
ns
***
Ct
40
30
20
10
0
GAPDH
POU5F1/OCT4
HNF4A
KLF4
DNMT3A
DNMT3B
DNMT1
MBD1
MBD2
MBD3
MBD4
MECP2
POLR2A
XRCC1
POLB
LIG3
APEX1
MYC
PARP1
SCD
SP1
PGAM5
KRAS
PTGS2
FBRS
JAK2
VEGFA
P4HA1
STAT3
DNAJC5
HEK293
TDGKO

Supplementary Figure 9. TDG knockout has no major effects on HEK293 cell biology. (A) Growth curve of HEK293 cells compared to HEK293 TDGKO cells. Circles represent individual biological replicates (n = 3) and lines connect mean values. There is no significant difference in cell number at any day as measured by independent t-test. (B-C) Total RNA (B) and DNA (C) yields from equal numbers of HEK293 and HEK293 TDGKO cells. There are no significant differences as measured by independent t-test. Circles represent values from n = 3 independent biological replicates and bars and error represent mean ± SEM. (D) Alignments of Sanger sequencing data of 6 selected significant 3XFLAG-TDG-N140A-CD peaks in the promoters of the indicated genes, showing no detectable mutations between HEK293 and HEK293 TDGKO cells. (E) RT-qPCR Ct values for indicated genes in HEK293 and HEK293 TDGKO cells. Raw Ct values are plotted to avoid normalization to a housekeeping gene in case of global differences in gene expression, though none were observed. Circles represent values from n = 3 independent biological replicates and bars and error represent mean ± SEM. HEK293 and HEK293 TDGKO Ct values were compared by multiple independent t-tests in GraphPad Prism v9.4.1. *** indicates $p<0.001$ after correction for multiple comparisons with the FDR method and ns indicates no statistically significant difference.



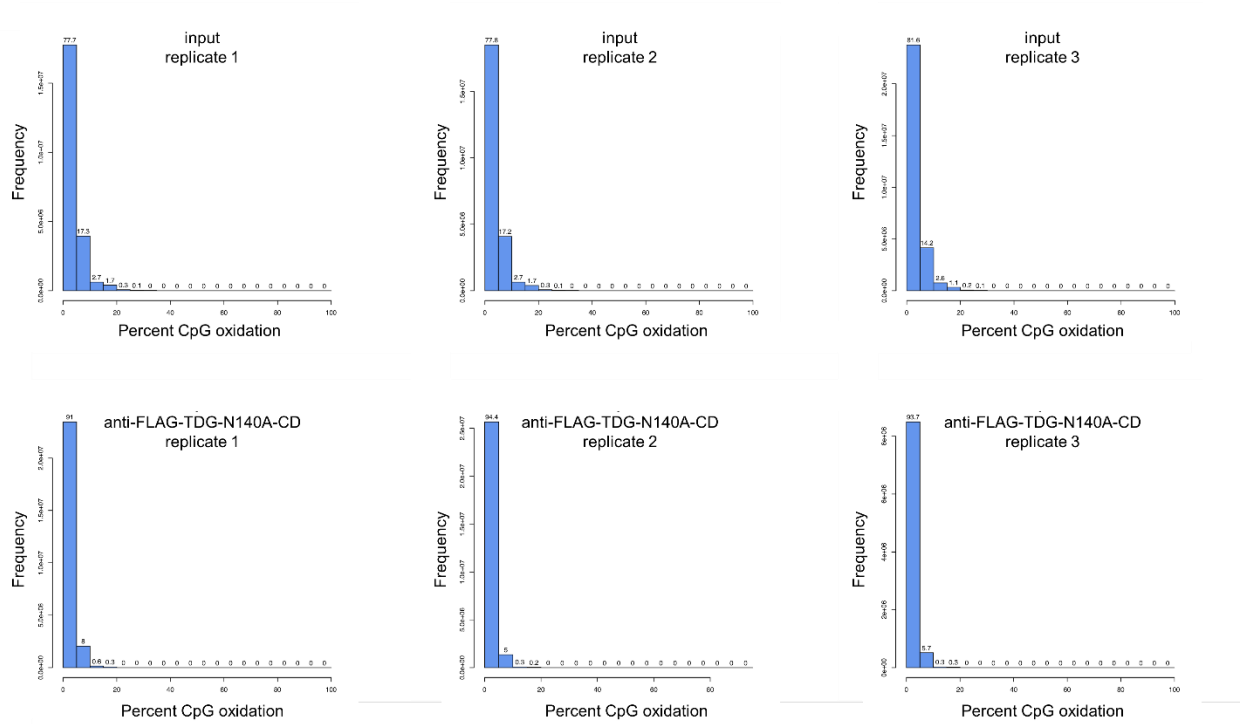
Supplementary Figure 10. The interaction between TDG and MBD3. (A) Interaction network of all statistically significantly enriched co-immunoprecipitated proteins in 3XFLAG-TDG-N140A-CD LC-MS/MS data, plotted with STRING-db. Each circle (node) represents a protein and is labeled with the protein name. Each line (edge) represents a line of evidence (compiled by STRING-db) for an interaction between the nodes that it links and is colored according to the legend. Nodes belonging to selected significantly enriched Reactome Pathways or Local network clusters (STRING) are colored according to the legend and described in greater detail in the Results. (B) Western blots of FLAG-tagged TDG co-immunoprecipitation (co-IP) experiments using 3X-FLAG-N140A-CD over-expressing HEK293 cells or control (empty vector) HEK293 cells, performed with 2 biological replicates, detected with anti-FLAG (top), anti-MBD3

(middle), or anti-MBD2 antibodies. Input samples represent total nuclear lysate and FLAG-IP samples were immunoprecipitated with anti-FLAG antibody. In FLAG-IP samples, IgG light and heavy chains from the mouse anti-FLAG antibody are visible due to the use of an anti-mouse secondary antibody, which are reduced in anti-MBD3 and anti-MBD2 blots which use primary antibodies produced in rabbit. An undefined large band in all samples from 3X-FLAG-TDG-N140A-CD cells is marked with an asterisk and corresponds to the predicted size of 3X-FLAG-TDG-N140A-2A-eGFP wherein the 2A peptide failed to induce ribosomal skipping. High exposure conditions are also included for better visualization of IP bands compared to more saturated input signal. A faint MBD3 band is visible in negative control co-IPs but MBD3 was not detected in LC-MS/MS data for negative controls. The marker is Froggabio Pink Plus Prestained Protein Ladder. (C) Same as (B) but performed in MCF-7 and HEPG2 cells with a single replicate and no MBD2 detection. “E” indicates empty vector as in (B). (D) Average ChIP-seq signal for TDG and MBD3 (this study) and MTA1, HDAC2, and GATAD2B (ENCODE data) binned over active genes (those with significant pol2 peaks in MCF-7 ENCODE data) and inactive genes (those without significant pol2 peaks), as well as ±20 kb from transcription start sites (TSS) and transcription end sites (TES). Unlike HEK293 data in Figure 5, for MCF-7, no RNA-seq data was used to identify high-confidence inactive genes and thus most inactive genes still exhibit pol2 binding that was below statistical significance thresholds, and thus binding at inactive genes is overestimated.

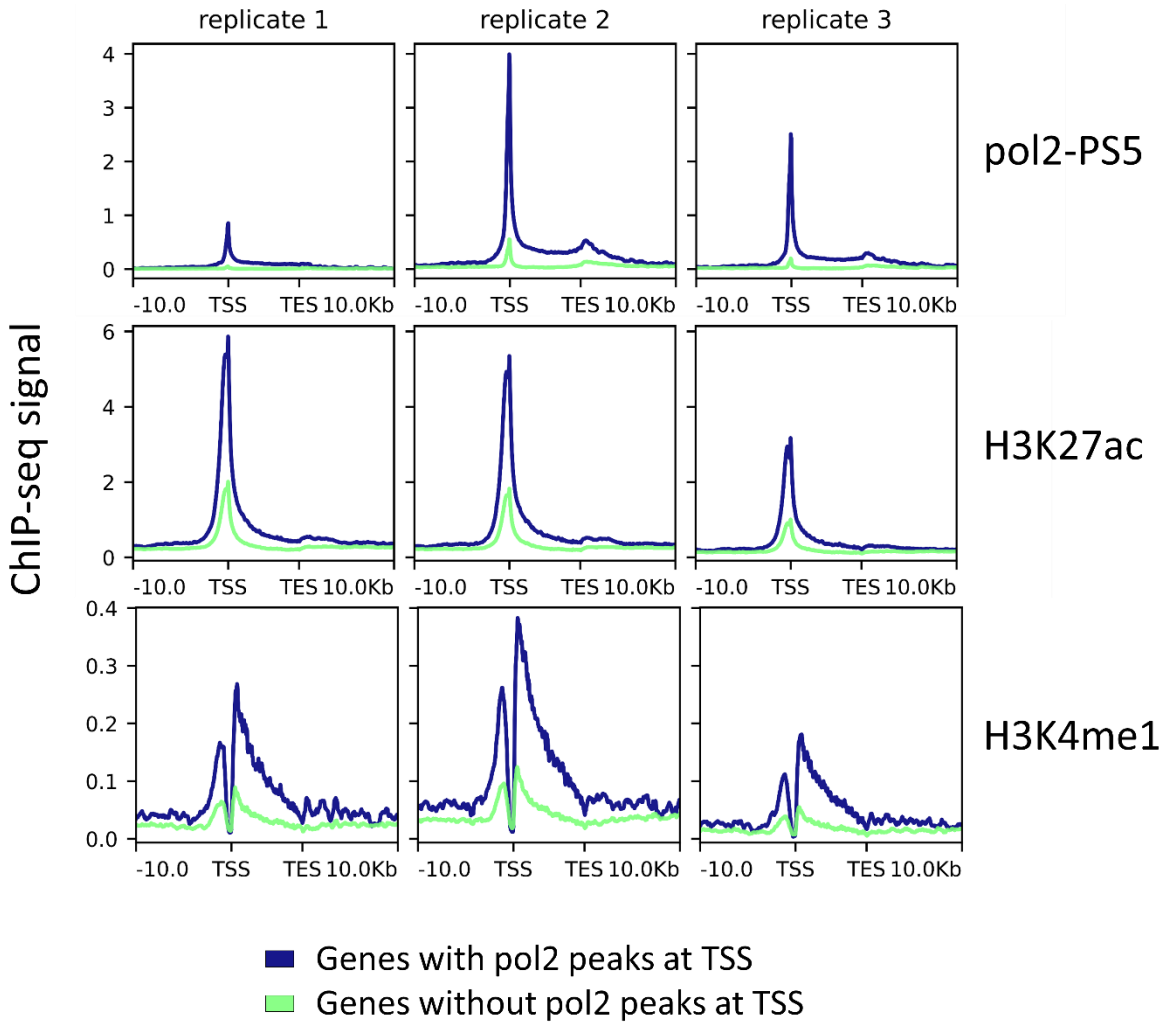

A
GAPDH
TDG binding
GAPDH in
GAPDH out
HEK293 control
MBD3 knockout
no FLAG control
B
RNA45SN1
RNA45S in
RNA45S out
C
RNA5S1
RNA5S in
RNA5S out
D
E
RNA45SN1
Input
FLAG-TDG ChIP
Percent Methylated Cytosines
CpG Number
p<0.0001
F
RNA45SN1
HEK293
TDGKO
Percent Oxidized Cytosines
G
RNA45SN1
HEK293
TDGKO
HEK293 MBD3 KO
TDGKO MBD3 KO
Percent Methylated Cytosines
H

Supplementary Figure 11. Gene expression regulation by TDG and MBD3. (A-C) ChIP-qPCR data, normalized to input samples, for ChIP performed with anti-FLAG antibody in HEK293 and MBD3 knockout HEK293 cells expressing 3XFLAG-TDG-N140A-CD or HEK293 cells expressing empty vector (no FLAG). qPCR was performed with primers specific for regions with significant 3XFLAG-TDG-N140A-CD peaks in HEK293 ChIP-seq data ("in") or adjacent regions without significant 3XFLAG-TDG-N140A-CD peaks ("out") in the *GAPDH* promoter (A), rDNA repeat promoter (*RNA45SN1*) (B), or a 5S rRNA promoter (*RNA5S1*) (C). Paired t-tests were used to compare "in" and "out" peaks within the same ChIP sample and unpaired t-tests were used to compare across ChIP samples. Data are presented as mean ± SEM with data for n = 3 independent biological replicates plotted as circles. (D) IGV genome browser screenshot depicting an rDNA promoter repeat with conserved MBD3 and TDG binding across cell lines. Replicated significant peaks of 3XFLAG-TDG-N140A-CD in HEK293 cells are shown in the top track, followed by significant 3XFLAG-TDG-N140A-CD peaks in each cell line used in this study (MCF7, IMR90, HEPG2) and significant MBD3 peaks in all cell lines used in this study as well as two replicate public datasets of MBD3 peaks in HEK293 cells. Light blue lines in plotted peaks represent peak summits. (E) DNA methylation levels of 26 CpGs in the rDNA repeat promoter determined by bisulfite-pyrosequencing in ChIP DNA bound to 3XFLAG-TDG-N140A-CD or its corresponding input (total DNA). The plotted p-value was calculated by a two-way ANOVA comparing input and ChIP samples across all CpGs and all CpGs were statistically differentially methylated between input and ChIP samples by unpaired independent t-test with correction for multiple testing by the FDR method, except for CpGs 2, 24, and 26 (not plotted). (F) Percent oxidized cytosines determined by APOBEC-pyrosequencing of the same CpGs as in (E) in control HEK293 cells and HEK293 TDGKO cells. (G) Percent DNA methylation of the same CpGs as in (F) in the rDNA promoter, measured by bisulfite-pyrosequencing in indicated cell lines. CpG methylation in HEK293 MBD3 KO and HEK293 TDGKO MBD3 KO were compared using unpaired t-tests with correction for multiple testing by the FDR method. (H) IGV genome browser screen shot depicting

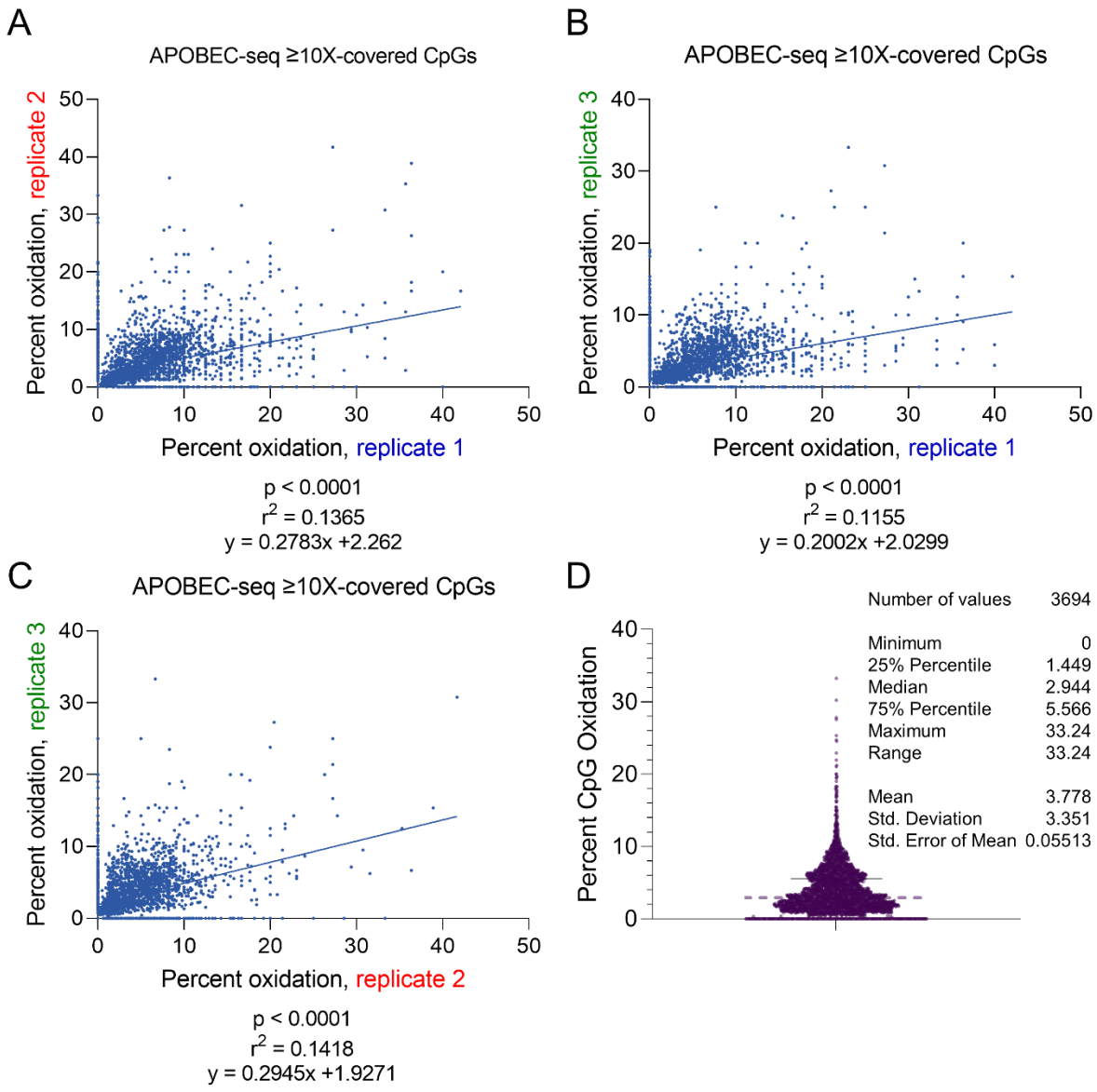
replicated significant 3XFLAG-TDG-N140A-CD peaks in a 5S rRNA gene cluster on chromosome 1 (as in **Figure 5**) but accompanied by significant MBD3 peaks in HEK293 cells, HEK293 TDG-N140A cells, and HEK293 TDGKO cells. All statistical tests were conducted in GraphPad Prism v9.4.1. * indicates $p<0.05$, ** indicates $p<0.01$, *** indicates $p<0.001$, and ns indicates no statistically significant difference.



Supplementary Figure 12. Histograms depicting percent CpG oxidation for each individual CpG covered by >10X reads in genome-wide APOBEC-seq split into 20 bins at 5% oxidation intervals.

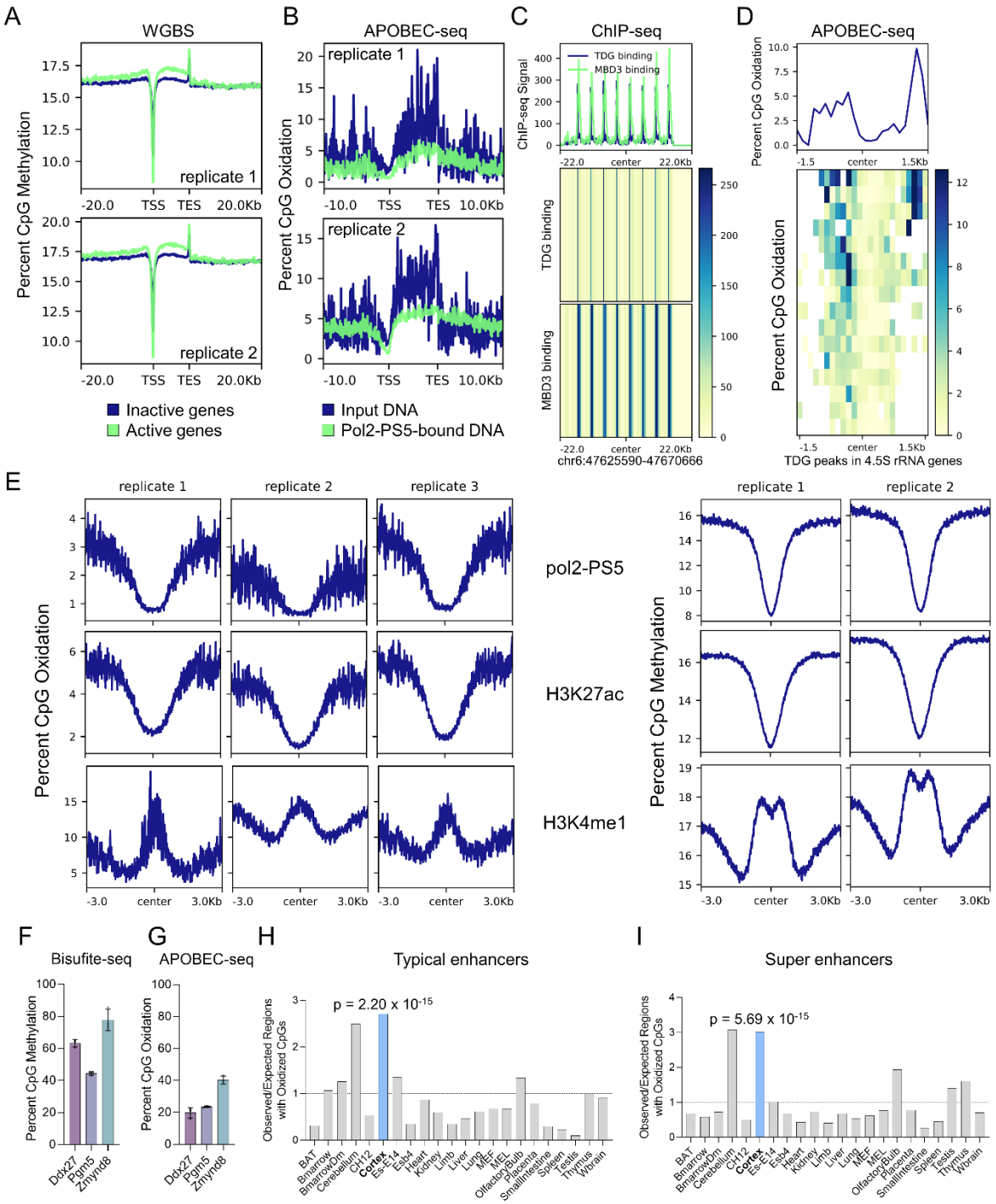


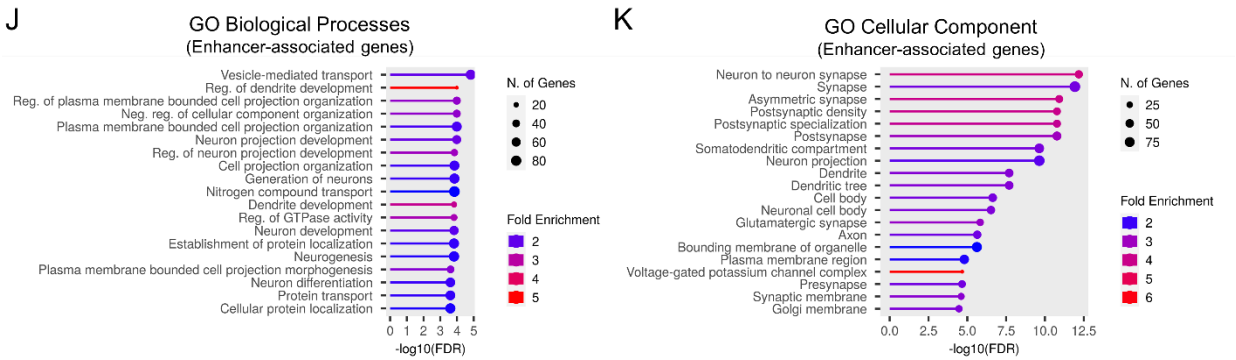
Supplementary Figure 13. Mouse cortex ChIP-seq data sets show typical binding profiles. ChIP-seq signal over gene structure in 3 mouse cortices for pol2-PS5 (top), H3K27ac (middle), and H3K4me1 (bottom) ChIP-seq data. Signal is separated for genes without significant pol2-PS5 peaks (green) and genes with significant pol2-PS5 peaks (blue) at the TSS. The region between TSS and TES is scaled to 10 kb.



Supplementary Figure 14. CpG oxidation rates are similar across samples. (A-C) XY scatter dot plots comparing percent CpG oxidation of each CpG between each pair of 3 replicate input APOBEC-seq datasets of adult mouse cortex. CpGs were only

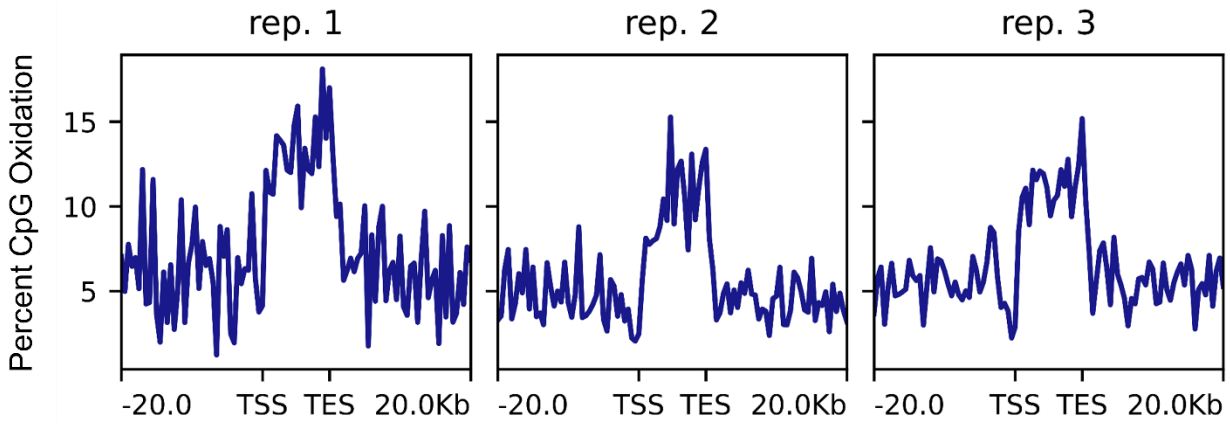
included if they were at least 10X covered in each of the three samples. The p-value, r^2 , and formula calculated with a simple linear regression are given below each plot. (D) Plot of percent oxidation of all CpGs from (A-C), averaged across the three samples. Line indicates 75% percentile and dashed line indicates median. All other descriptive statistics are shown.


A
WGBS
Percent CpG Methylation
replicate 1
replicate 2
-20.0 TSS TES 20.0Kb
Inactive genes
Active genes
B
APOBEC-seq
Percent CpG Oxidation
replicate 1
replicate 2
-10.0 TSS TES 10.0Kb
Input DNA
Pol2-PS5-bound DNA
C
ChIP-seq
ChIP-seq Signal
TDG binding
MBD3 binding
-22.0 center 22.0Kb
chr6:47625590-47670666
D
APOBEC-seq
Percent CpG Oxidation
-1.5 center 1.5Kb
TDG peaks in 4.5S rRNA genes
E
replicate 1
replicate 2
replicate 3
Percent CpG Oxidation
pol2-PS5
H3K27ac
H3K4me1
Percent CpG Methylation
-3.0 center 3.0Kb
F
Bisufite-seq
Percent CpG Methylation
Ddx27
Pgm5
Zmynd8
G
APOBEC-seq
Percent CpG Oxidation
H
Typical enhancers
p = 2.20 x 10-15
Observed/Expected Regions with Oxidized CpGs
I
Super enhancers
p = 5.69 x 10-15
BAT
Bmarrow
BmarrowDm
Cerebellum
CH12
Cortex
Es-E14
Es-E4
Heart
Kidney
Limb
Liver
Lung
MEF
MEL
OlfactoryBulb
Placenta
Smallintestine
Spleen
Testis
Thymus
Wbrain

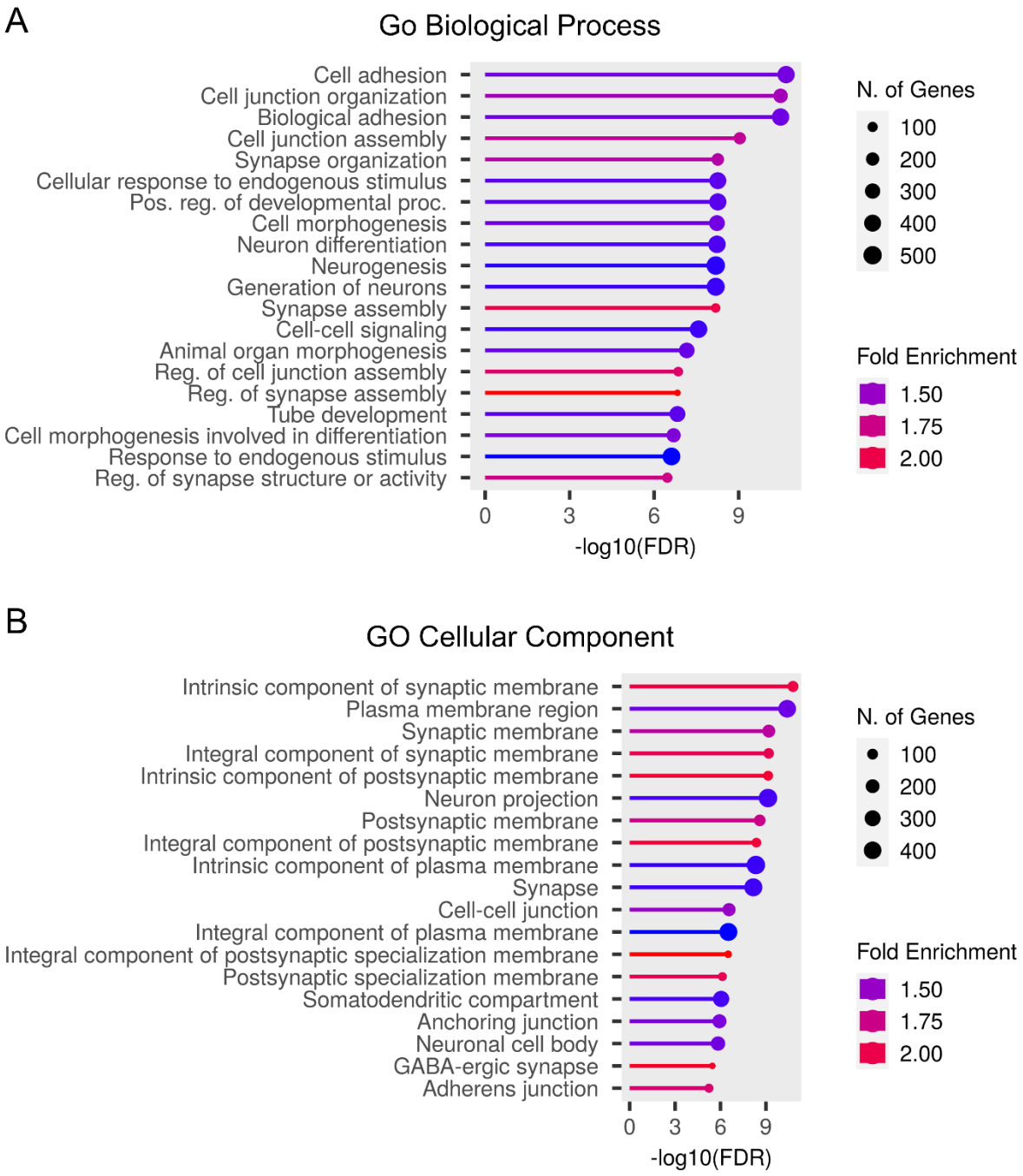


Supplementary Figure 15. Extended analysis of CpG oxidation in the adult mouse cortex. (A) Percent CpG methylation averaged over gene structures, plotted as a function of distance from the transcription start site (TSS) or transcription end site (TES), for CpGs in public mouse cortex WGBS data with a minimum of 10X coverage in each of 2 replicates, separated across genes with significant pol2-PS5 binding at their promoters (green, active genes) and those without (blue, inactive genes) based on data generated in this study. The region between TSS and TES is scaled to 10 kb. (B) Percent CpG oxidation averaged across all gene structures in input DNA (blue) or pol2-PS5-bound DNA (green) for two replicates, using a minimum of 10X coverage for APOBEC-seq. (C) ChIP-seq signal of TDG (blue) and (MBD3) green showing overlapping binding and broader MBD3 peaks in rRNA 4.5S genes on chromosome 6. (D) Average CpG oxidation measured by APOBEC-seq (>10X coverage) binned across 3 kb of 16 peaks of dual MBD3/TDG binding in the mouse cortex, where the center indicates peak center. Bottom panel displays average CpG oxidation of each bin across each individual peak. White bins indicate no data. (E) Average CpG oxidation (left panels) or CpG methylation (right panels) as a function of distance from all significant peaks of pol2-PS5 (top), H3K27ac (middle), or H3K4me1 (bottom) binding, centered at the peak center, supported by CpGs with >10X coverage for each of 3 APOBEC-seq replicates and 2 WGBS replicates. APOBEC-seq data is from bound DNA in each ChIP experiment. (F-G) Validation of 3 individual CpGs in the promoters of the indicated genes by bisulfite-pyrosequencing (F) or APOBEC-pyrosequencing (G). Each circle represents a cortex from an independent animal (n = 3), the bar plots the mean value,

and error bars represent SD. Each gene contains a single CpG. (H-I) Observed/expected values of the number of (tissue-specific) enhancers containing high-confidence oxidized CpGs from each indicated tissue type for typical enhancers (H) and super enhancers (I). Consult Sethi et al. for enhancer, tissue, and identification methods descriptions. (BAT: brown adipose tissue; Bmarrrow: bone marrow; BmarrowDm: bone marrow derived macrophage; CH12: B-cell lymphoma; Esb4: mouse embryonic stem cells; Es-E14: mouse embryonic stem cell line embryonic day 14.5; MEF: mouse embryonic fibroblast; MEL: leukemia; Wbrain: whole brain). (J-K) Gene ontology term enrichment analysis for list of genes functionally associated with enhancers which contain high-confidence oxidized CpGs, using GO Biological Processes (J) or GO Cellular Component (K) terms.

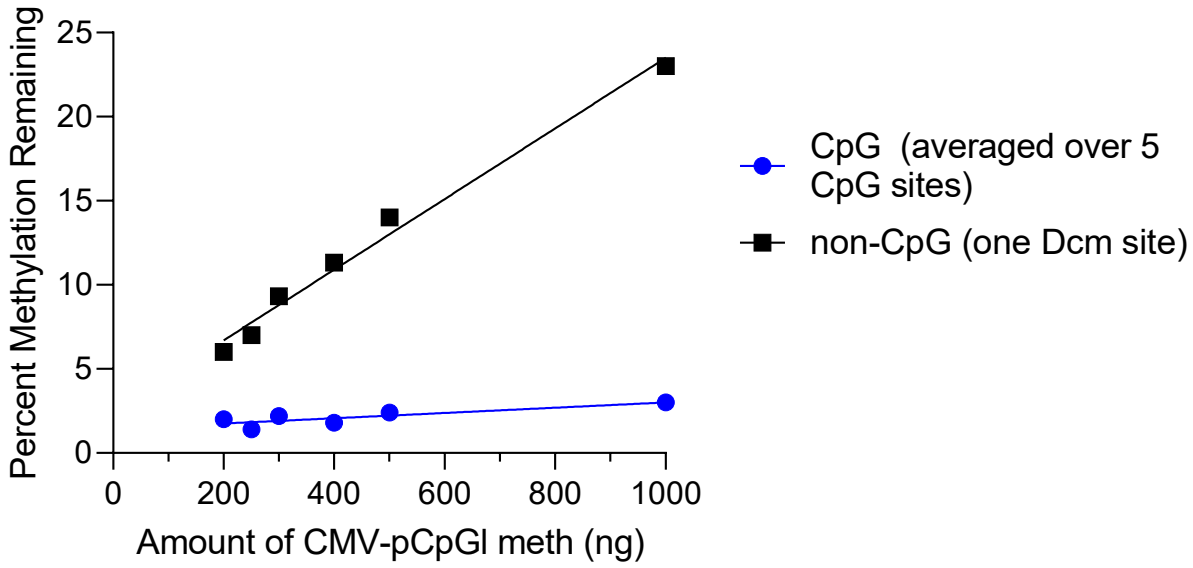


Supplementary Figure 16. Genes with high-confidence oxidized CpGs show primarily gene body oxidation. Average oxidation percentages of all high-confidence oxidized CpGs that occurred near genes in three cortices subjected to APOBEC-seq, plotted over gene structures from the transcription start site (TSS) to the transcription end site (TES) and an additional 20 kb in each direction.



Supplementary Figure 17. Genes with methylated CpGs are not as neuron-specific as genes with oxidized CpGs. (A-B) Gene ontology term enrichment analysis

for list of genes with methylated CpGs in the mouse cortex from public WGBS data using GO Biological Processes (A) or GO Cellular Component (B) terms.



Supplementary Figure 18. Optimization of maximum methylated CMV-pCpGl quantity for sufficient oxidation by NEB TET2. Percent methylation remaining after TET2 reaction from an originally fully methylated plasmid, measured by bisulfite-pyrosequencing, as a function of CMV-pCpGl quantity for one non-CpG site (Dcm, black) or averaged across the 5 CpGs in the CMV pyrosequencing assay (blue). The plotted value is the average of three technical replicates. The final reaction volume was 50 µL and all other reaction components were kept the same (except reduced water to account for increased DNA volume) as described in the Methods.

Antibody	Animal no.	IP yield (ng)	Mapped reads	Cs analyzed	CpG oxidation	CHG oxidation	CHH oxidation	CN oxidation	Conversion rate
None - input	5		12013243	382011847	13.4	2.6	3.1	3.2	98.94%
None - input	7		13552067	446563326	10.3	1.8	2	2.2	99.27%
None - input	9		17698680	604648817	10.1	1.7	1.9	2	99.48%
H3K4me1	5	821.6	19637092	739909705	11.8	1.5	1.6	1.8	99.32%
H3K4me1	7	785.2	42809121	1615097750	11.8	1.7	1.8	2	99.03%
H3K4me1	9	889.2	17938748	687560663	11.7	1.7	1.8	2	99.28%
H3K27ac	5	187.2	15615515	636758970	5.3	1.3	1.4	1.7	99.49%
H3K27ac	7	133.64	16186925	693502249	3.6	1	1.1	1.7	99.56%
H3K27ac	9	208.52	16720284	702601635	4.6	1	1.1	1.6	99.61%
POL2-PS5	5	10.61	20731853	989132113	2.8	0.5	0.5	1.3	99.64%
POL2-PS5	7	8.27	10469890	505640226	2.8	0.6	0.7	2.2	99.62%
POL2-PS5	9	9.46	19980588	966958730	2.9	0.5	0.6	1.7	99.61%

Supplementary Table 1. APOBEC-seq summary statistics in mouse cortex.

Application	Antibody	Vendor	Catalog No.	Dilution for western blot
Western blot	TDG	Abcam	ab154192	1 in 2,000
Western blot	FLAG	Millipore Sigma	F1804	1 in 2,000
Western blot	Beta actin	Abcam	ab8227	1 in 5,000
Western blot	APEX1	Abcam	ab189474	1 in 1,000
Western blot	MBD1	Abcam	ab108510	1 in 500
Western blot	MBD2	Abcam	ab188474	1 in 1,000
Western blot	MBD3	Abcam	ab188401	1 in 5,000
Western blot	MECP2	Abcam	ab253197	1 in 1,000
Western blot	Goat Anti-Rabbit IgG H&L	Abcam	ab6721	1 in 10,000
Western blot	Rabbit Anti-Mouse IgG H&L	Abcam	ab6728	1 in 10,000
ChIP	TDG	Proteintech	13370-1-AP	
ChIP	MBD2	Abcam	ab188474	
ChIP	TBP	Abcam	ab220788	
ChIP	MECP2	Abcam	ab253197	
ChIP	MBD3	Abcam	ab157464	
ChIP	MBD4	Abcam	ab224809	
ChIP	MBD1	Abcam	ab108510	
ChIP	H3K27ac	Abcam	ab4729	
ChIP	H3K4me1	Abcam	ab176877	
ChIP	pol2-PS5	Abcam	ab5408	
ChIP	FLAG	Millipore Sigma	F1804	

Supplementary Table 2. All antibodies used in this study.

Summary of the thesis and contributions to original knowledge

The major motivation for the work in this thesis was to address general gaps of knowledge in the DNA methylation research field and to build technologies that not only facilitated these endeavours but also can continue to aid researchers throughout the field. We identified these general gaps to be the inability to reliably attribute causality (in gene expression regulation) to any specific instance of DNA methylation – which was addressed in Chapter 2 – and the general lack of detailed insight into the active DNA methylation pathway, which was addressed in Chapter 3, primarily in the context of transfected plasmids. To this end, this thesis contributes the following original knowledge in Chapter 2:

1. In a series of experiments, I showed that dCas9-TET approaches are confounded by off-target effects and non-catalytic activities that do not allow them to be used to assess the causal role of DNA demethylation events in gene expression regulation.
2. I demonstrated a novel application of CRISPR/dCas9 technology in showing that a nuclease-dead dCas9 could prevent the activity of methyltransferases at specific target sites.
3. In order to increase utility to other researchers in targeting specific CpGs, I determined the size and characteristics of the DNA region that is protected from methylation.
4. I showed that this method can be used in dividing cells, optimized it for near complete demethylation of targeted sites, and presented a system by which dCas9 could be removed after demethylation, such that the effect of this demethylation could then be assessed.
5. I reported the transcriptional responses of several endogenous genes to DNA demethylation using this steric interference method and showed that DNA demethylation at many sites does not affect expression, whereas at other sites in can modulate minor changes in expression.

6. I showed that only DNA demethylation of the CGG repeat in Fragile X syndrome patient fibroblasts can re-activate expression of the silenced *FMR1* gene.
7. I performed extensive off-target analysis to demonstrate that this method is specific to the targeted site, unlike dCas9-TET methods.
8. I demonstrated that DNA demethylation can accompany CRISPR gene editing experiments and suggested that the consequences of this activity should be considered in gene editing experiments.
9. Overall, I showed that demethylation with dCas9 alone is an efficient and simple method for researchers to use to assess the causal impact of DNA demethylation at specific sites on gene expression.

In Chapter 3 of this thesis, I made the following discoveries that stand as contributions to original knowledge:

1. APOBEC-seq is a novel nondestructive sequencing technique for the specific detection of oxidized methyl-cytosines which we show to be simple to use, sensitive, specific, and applicable genome-wide.
2. Oxidized promoters are expressed in all tested human cell lines.
3. Expression of oxidized promoters requires the activity of TDG.
4. Catalytic TDG mutants can be used for enrichment of oxidized DNA.
5. TDG-mediated demethylation parallels re-activation, but TDG has transcriptional activation activity that confounds the notion that demethylation is required for re-activation.
6. Transcription is not required for demethylation by TDG.
7. The base excision repair enzyme APEX1 is not required for the efficient demethylation by TDG.
8. Full-length TDG, but not its catalytic domain (which is sufficient for highly specific oxidized CpG binding) is required for re-activation of oxidized promoters.

9. The MBD family of proteins all exhibit binding to oxidized promoters in addition to methylated, while MBD3 is specific only for oxidized.
10. TDG binds all active promoters in HEK293 cells, regardless of whether they are transcribed by RNA polymerase I, II, or III, suggesting it safeguards active promoters from methylation/oxidation.
11. TDG interacts with the MBD3/NuRD complex, which is partly responsible for the recruitment of TDG to active promoters.
12. TDG binding sites in HEK293 cells are not oxidized.
13. Oxidized CpGs are depleted from promoters and nearby 5' UTRs, where TDG is most enriched, and instead are significantly enriched in regions marked with H3K4me1 and H3K9me3.
14. Oxidized CpGs in the mouse brain are also depleted from promoters, where TDG is found, but instead occur in gene bodies where TDG is absent, as well as in H3K4me1 sites. Genes and enhancers with oxidized CpGs are highly-tissue specific.

Chapter 4: General Discussion

Detailed discussions of experimental results can be found in the respective Discussion sections of each Chapter that makes up the body of the thesis. However, there remain some broader points worthy of further discussion.

16.1.1 Applications of dCas9-based demethylation in dividing cells

Here, I developed an efficient targeted DNA demethylation method based on the principle of interference with DNMT1 at a specific site by a targeted dCas9 protein and, not included in this thesis, I further published an optimized protocol to help DNA methylation researchers to implement this new method more efficiently[3]. DNA demethylation by steric interference of DNMT1 using CRISPR/dCas9 as described herein can be applied to study DNA methylation of specific CpGs in any dividing cell line as it relates to any physiological or pathological condition. The simplest application of this method is to interrogate the causal role of methylation of single CpGs – or different combinations of specific CpGs – in the regulation of gene expression. The association between methylation levels of single CpGs and gene expression has been demonstrated in several clinically significant genes: some examples are *TP53*[532], *KIT*[533], *ESR1*[534], *IL6*[535], and *TLR2*[536]. However, across all such correlative examples, there is no causal evidence that DNA methylation changes at a single cytosine lead to changes in gene expression in their endogenous cellular contexts. Other targeted DNA methylation editing methods (TET or DNMT fusions to zinc-fingers, TALEs, or dCas9) are typically not suited to modify single CpG as the enzymatic domains are flexibly tethered and affect methylation of cytosines across large genetic regions[311]. This work presents a new way to address this question. By using this steric hindrance method, researchers can focus on the impact of specific sites and avoid the pitfalls that accompany the broad, off-target, and non-catalytic effects of dCas9-tethered epigenetic enzymes. The transitory nature of steric inhibition (i.e., by loss of dCas9 expression) also allows researchers to study the effects of loss of DNA methylation at defined time

points during differentiation, aging, and other physiological processes on downstream events and methylation trajectories.

One key consideration for these experiments is that demethylation would be expected to exert varying consequences on gene expression in different tissues or cell lines, depending on the transcription factors that are expressed in that cell line. For example, if the purpose of the experiment is to assess the impact on gene expression of a target CpG that is differentially methylated in the brain tissue of a rat model of a psychiatric disorder, the highest likelihood of achieving differential gene expression consequent to targeted DNA demethylation would most probably be in an appropriate rat neuronal cell line where the relevant transcription factors required to activate the gene are likely to be expressed: if the potential interactors of a target CpG are not expressed in the cell line under study, it can be expected that any changes in the methylation state of the target CpG would not lead to functional differences.

Another logical extension of this method is in preventing methylation at rarer non-CpG[537] and, potentially, non-cytosine[10] sites. Non-CpG cytosine methylation is a poorly understood phenomenon that has been shown to exert consequences on gene expression and development[538,539]. Establishment and potential maintenance of non-CpG methylation patterns is thought to be dependent on DNMT3A and DNMT3B complexed with DNMT3L[61,537,538] and interference with this activity by a modified dCas9-demethylation protocol is poised to reveal new insights into the biological roles of non-CpG methylation.

16.1.2 Potential clinical applications of dCas9-based demethylation

Steric hindrance cannot induce loss of nondynamic DNA methylation in non-dividing, terminally differentiated cells such as neurons or muscle cells. This limitation restricts its utility in addressing the role of site-specific methylation in fully differentiated systems and limits its clinical use. However, an agent that can cause site-specific demethylation

in dividing cells still has a wide range of utility. One application is in the use of ex vivo methods with either patient-derived cells, induced pluripotent cells, or progenitor cells that can be epigenetically modified and (re)introduced to the patient. For example, targeted demethylation could be implemented as a supplemental modification of otherwise engineered T-cells as part of chimeric antigen receptor (CAR) T-cell therapy by demethylating genes that augment T-cell anticancer activity and thus increase overall therapeutic efficacy[540,541]. Another example would be in the targeted demethylation of the promoters of the insulin gene and of other pancreatic transcription factors in order to increase the efficiency of approaches aimed to treat diabetes by the transdifferentiation of liver cells into insulin-secreting pancreas-like cells[542]. Beyond ex vivo applications, one potential implementation of dCas9-based demethylation is in the context of cancer cells. Rapid cell division and epigenetic dysregulation are hallmarks of cancer thus the technique is both mechanistically feasible and therapeutically relevent. Site-specific demethylating agents can potentially demethylate and activate tumor suppressor genes while avoiding demethylation of tumor-promoting and metastatic genes to improve anticancer therapy.

As the use of DNA methylation inhibitors in the treatment of various clinical conditions is becoming increasingly supported, the need for site-specific demethylation agents that can be administered in a clinical setting is growing and represents the founding principle of several biotechnology companies. There are two major obstacles currently delaying the clinical utility of targeted epigenetic engineering technologies: (1) difficulties in the delivery of CRISPR/Cas9 components to target cells and (2) molecular technologies that fail to achieve specific and efficient epigenetic editing. While the targeted steric hindrance protocols developed herein may address the latter issue to potentially reduce undesired side effects, their advancement as clinical tools would stand to benefit from the active research efforts into the optimization of the delivery of CRISPR/Cas9 components in DNA, RNA, or ribonucleoprotein form. The above emphasis on the utility of this method in ex vivo settings is based on the fact that CRISPR/dCas9 delivery to

isolated cells is currently feasible by numerous technologies – including, most commonly, electroporation[543-545], but also by physical injection, lipid-based transfection reagents, viral vectors, and lipid nanoparticles – and, with few exceptions, ex vivo therapy is nearly the sole paradigm across several dozen CRISPR/Cas9-based therapies currently undergoing clinical trials[545]. Targeting CRISPR/Cas9 components to specific organs *in vivo* is inherently more difficult. Potential solutions include local adeno-associated virus (AAV) injection to tissues such as the retina[546] or brain[547] or the development of lipid nanoparticles[548-551] or peptides[552,553] with tropisms for specific organs or with tissue-restricted activities[549] that can deliver CRISPR/Cas9 components as ribonucleoproteins. An interesting facet of the dCas9-based demethylation mechanism is that it involves transient occupancy of the binding site – long enough to cause demethylation – but not persistent occupancy: it must be removed for the demethylated site to become accesible to other nuclear factors. This complements therapeutic applications, as CRISPR/Cas9 components delivered as drugs will degrade with time. The kinetics of delivery of dCas9 protein-gRNA complexes or dCas9-gRNA RNA could be optimized to achieve the highest blockage of DNA methylation before the concentration of the complexes is reduced and interaction with the transcription machinery is enabled. Delivery is currently a general challenge in the entire CRISPR/Cas9 and gene-editing field, and its potential resolution as a result of global efforts could turn steric hindrance into a potent pharmacological, site-specific demethylation agent.

16.1.3 Expanding applications of dCas9-based demethylation beyond dividing cells

Across this work, the successful application of this method is, in principle, dependent on the DNMT1 activity in dividing cells and is thus only feasible in dividing cells. Therefore, why is the potential improvement of delivery to differentiated, largely nondividing organs *in vivo* an exciting prospect in the context of dCas9-based demethylation? Dynamic DNA methylation systems in non-dividing cells, such as neurons, involve both de novo

methylation and demethylation events in the absence of DNA replication. De novo methylation is catalyzed by de novo methyltransferases like DNMT3A[41,50,63]. It is therefore possible to prevent de novo methylation in response to, for example, exposure or learning and memory episodes by treating with site-specific dCas9-gRNA before the anticipated trigger. This treatment would sterically inhibit future de novo methylation. Removal of dCas9 or its turnover would then enable interaction between the unmethylated position and the transcription machinery and other factors in the future. Such an approach could help understand the role of site-specific de novo methylation in neuronal activation, stress responses, and other context-dependent processes.

Dynamic DNA methylation responses, such as those observed during neuronal activation, involve not only de novo methylation but also site-specific demethylation of specific CG positions in key genes[231,554]. Since site-specific demethylation relies on interactions between TET enzymes and DNA, it is also potentially amenable to inhibition by this steric hindrance method. By applying steric hindrance prior to neuronal activation or other triggers, demethylation at the targeted CG site could similarly be prevented. Upon removal of dCas9 or its turnover, the methylated site could then interact with other factors once the transient demethylation trigger subsides. This would allow for the assessment of the functional role of site-specific demethylation by comparing the physiological and phenotypic responses of animals that retain methylation at the position and those that were demethylated in response to treatment. Although the inhibition of the binding of TET enzymes by gRNA-targeted dCas9 has not yet been tested, it is a promising approach that fundamentally parallels DNMT inhibition and could be optimized for studying site-specific demethylation in dynamic systems.

16.1.4 Limitations of dCas9-based demethylation

The precise physical distance over which dCas9 interferes with mammalian DNMT binding has not yet been determined. However, there are some insights from dCas9 inhibition of the bacterial methyltransferase M.SssI *in vitro* in Chapter 2. Strict guidelines

for DNA demethylation success are to include the target CpG motif directly within the gRNA sequence or the NGG protospacer adjacent motif (PAM) sequence. Still, based on the data presented in Chapter 2, it is reasonable to expect that steric interference of DNMT will occur as far as 5 base pairs upstream of the 5' end of the gRNA and 10 base pairs downstream of the 3' end.

It is clear from the mechanism of steric hindrance by dCas9 that a single CpG within a set of adjacent or nearby CpGs (e.g., 5'-CGCG-3') would be highly difficult to demethylate without affecting the other CpG. However, it is not impossible, as dCas9 could potentially be targeted such that the boundary of the steric interference falls between the two adjacent CpGs. This may be difficult to achieve, especially if there are still more adjacent CpGs, but could also be resolved by refining dCas9 positioning using other dCas9 orthologs with different physical sizes and PAM requirements[555].
On the other hand, if large CpG-rich regions are to be demethylated, the limited size of steric interference of dCas9 could become an issue. One possible resolution to this is the co-expression of multiple gRNAs through multiplexing[556,557] to target demethylation to a larger region.

Another difficult target for the dCas9-demethylation system is represented by highly methylated CpGs located within dense heterochromatin, which is, almost by definition, enriched in highly methylated CpG targets[558]. Heterochromatin regions are well-established as more difficult to target by CRISPR/Cas9-based systems due to the compact and inaccessible nature of the DNA[559-563]. Interestingly, there is evidence that TALE-based systems may be several-fold more efficient in targeting heterochromatin than CRISPR/Cas9-based systems[564]. Therefore, a matter for future investigation is whether TALE- and zinc-finger-based interference with the DNA methyltransferase machinery could mimic and possibly augment the demethylation capacity of the CRISPR/dCas9 system described herein, particularly for CpG-dense regions.

Demethylation is also highly dependent on efficient on-target dCas9 binding to outcompete the binding of endogenous DNA methyltransferase. This is particularly important in the case of DNMT1, which is typically highly expressed[565] and associates with the replication fork[45], presenting the potential risk of replication-associated helicase activity to displace target-bound dCas9[566] and allow nearby DNMT1 to access the target site more rapidly than dCas9 can reassociate with it. Therefore, a high stochiometric ratio of dCas9 and gRNA compared to DNA methyltransferase is critical for efficient DNA demethylation. While we have demonstrated this to be feasible in multiple cell lines, high and lasting dCas9 and gRNA expression levels were achieved with lentiviral components: this may be a challenge in cells – whether in culture or *in vivo* – that are more difficult transduce[567], in which the promoters driving dCas9 and gRNA expression (in these experiments, human ubiquitin C and human U6 promoters, respectively) might be less active[568], or when other, clinically relevant, non-viral modes of delivery are used. However, we did provide evidence that transfection of dCas9 and gRNA plasmids into mouse cells can produce significant targeted demethylation. This proof-of-concept suggests that non-viral delivery methods for transient expression could be used for dCas9-based demethylation and represents a promising basis for future work.

16.1.5 Compatibility with evolving CRISPR technologies

Due to the simple mechanism of steric interference by a catalytically inactive Cas9 protein, the dCas9-based demethylation approach would continue to benefit from the rapidly evolving CRISPR/Cas technology. While the method presented herein has been shown to cause complete demethylation, it relies on relatively early CRISPR/Cas9 technology using *S. pyogenes* dCas9 protein and standard 20-bp guide RNA. One point for modification is the *S. pyogenes* dCas9; it can be replaced with one of the dozens of dCas9 orthologs[555] that have since been discovered, particularly for the purpose of targeting a region where no 5'-NGG-3' *S. pyogenes* PAM is available, or for higher-fidelity editing by synthetically optimized dCas9 proteins[569-571]. In addition to the theoretical rationale which suggests that any catalytically inactive Cas9 ortholog could

sterically interfere with DNMT1 and cause demethylation, we have experimentally demonstrated the demethylation potential of at least one additional dCas9 protein from *S. aureus.* Similarly, improvements are expected from the continued evolution of gRNA design and users may invoke more recent advances such as the use of shorter gRNA lengths[572] or more accurate gRNA design algorithms[573].

16.2.1 Evidence for a potential DNA demethylation complex at active promoters

In Chapter 3, I described a sharp peak of TDG binding at active transcription start sites. Though a general increased binding of TDG to promoters has been previously reported[574], the work in this thesis reveals a surprisingly discrete and ubiquitous binding activity to nearly all active promoters, regardless of whether promoter expression is regulated by RNA polymerase I, II, or III. This observation, combined with a similar presence of MBD3/NuRD and TET enzymes at active TSS, suggests the potential presence of a DNA demethylation complex at active promoters that may work to prevent abberant hypermethylation of these active promoters. Moreover, the fact that such ubiquitous binding can be observed in HEK293 cells – which do not have detectable levels of oxidized 5mC derivatives in their promoters – implies that this machinery is not recruited as a response to methylation or oxidation but is instead consistently present at active promoters. The notion that unmethylated state of active promoters is actively maintained – or, at least, reinforced by such a complex – would represent a shift in our understanding of how the genomic DNA methylation pattern is regulated. Interestingly, a recent landmark study reported the presence of a similar oxidative damage (8-oxoguanine) repair complex that specifically binds to and safeguards promoters and enhancers from oxidative damage in advance of any damage[575]. Though 8-oxoguanine is distinct from oxidized 5mC derivatives, many of the same repair factors, such as APEX1 and PARP1, are also recruited as part of this oxidative repair complex[575]. The parallels between these findings and those reported in this thesis – a ubiquitous but specific preexisting protection of critical regulatory elements – might represent a set of

partially overlapping safeguards that exist as part of a distinct cellular pressure to ensure that promoters remain unmethylated, unoxidized, and undamaged.

16.2.2 Blurring the boundaries between repair and epigenetics: a more integrated view of the human regulatory network

A major potion of Chapter 3 aimed to disentangle transcriptional activation from demethylation as well as the presence of oxidized 5mC derivatives from a replacement with unmethylated cytosine. While these were important manipulations that are required to understand the individual contributions of each process to gene expression, there is an argument to be made that such distinctions are largely semantic and irrelevant to cellular physiology. If a multifunctional enzyme (e.g., TDG) is recruited to a genomic position as a result of CpG oxidation, does it matter which of its activities results in gene re-activation? Despite the fact that DNA repair and epigenetics represent two separate scientific disciplines, DNA modifications can be described from both perspectives. 8-oxoguanine has long been the classic example of DNA damage as a result of oxidative stress[576], represents the most commonly oxidized base in human DNA[577], and its repair mimics the active DNA demethylation pathway wherein glycosylation is performed by OGG1 (rather than TDG) and the site is then processed by BER in a manner that is likely identical to the BER steps of active DNA demethylation[577]. Yet, recently, multiple lines of evidence have suggested a regulatory role of 8-oxoguanine, marked by the observation that this modification and its repair lead to upregulation of gene expression[578-580]. At the same time, like 8-oxoguanine, cytosine methylation leads to increased mutation rates – most likely because of spontaneous deamination of 5mC to thymine[581] – and therefore 5mC and its derivatives can also be interpreted as a form of DNA damage. Both 8-oxoguanine and 5mC represent small nucleobase modifications in the major groove of DNA and thus both could affect transcription factor binding[577]. Moreover, both OGG1[580] and TDG (as shown in this thesis) exhibit transcriptional activation activity. These observations simultaneously highlight and obscure the research questions that were aimed to be answered by the work presented in this

thesis: it is important to remember that the active DNA demethylation process, though named for just a single consequence, does not only involved DNA demethylation but is an entire pathway involving numerous other changes, including those mediated by glycosylase transcriptional activation activity, recruitment of transcription factors by BER machinery, and modified transcription factor binding by oxidized 5mC derivatives and, seperately, by unmethylated cytosine. So, perhaps the most apt future approach to understanding the active DNA demethylation pathway is not one in which arbitrary boundaries are declared between these dependent processes, but one that considers the promiscuous activities of these multifunctional components and the multiple outputs of this process.

16.2.3 Components of the active DNA demethylation pathway: similarities between current findings and the evolving literature

Another focus of Chapter 3 was to better study the components of the active DNA demethylation pathway, as the current understanding of the protein members of this pathway is largely based on a single study that reconstituted the pathway *in vitro*[212]. In doing so, I reported two findings that are not necessarily completely novel, but are both controversial. A physical interaction between TDG and MBD3/NuRD that links MBD3/NuRD to the active DNA demethylation pathway has not been previously reported. Interestingly though, one study identified the MBD3/NuRD complex as a specific binder of 5fC-containing DNA[262]. This is a fascinating observation given that our study was performed in HEK293 cells with no apparent 5fC, but these cells exhibited an interaction between TDG and MBD3/NuRD and MBD3/NuRD was responsible for the recruitment of TDG to active unmethylated promoters. The recognition of 5hmC by MBD3 has been directly demonstrated[263] and also directly refuted by numerous groups[262,364], making it a controversial finding. However, this thesis also demonstrates the first evidence of a specific binding activity of MBD3 to highly oxidized DNA, suggesting that the link between MBD3, TDG, and active DNA demethylation might be cooperatively mediated by each of these two novel findings: recruitment of MBD3 to

oxidized DNA and its interaction with TDG. The convergence of these two independent studies on a possible relationship between MBD3/NuRD and components of the active DNA demethylation pathway – albeit by two different mechanisms – suggests that this may be indeed be a bona fide relationship and could possibly explain the important roles of MBD3/NuRD in pluripotency and development[494,519].

The other potentially controversial finding reported in this thesis is the fact that APEX1 is not required for TDG-triggered replacement of oxidized cytosines with unmethylated cytosines. As it has been reported that APEX1 knockout fully abrogates BER activity in the context of an AP-site-containing substrate in HEK293 cells[478], we were surprised to find that the BER pathway still was functional in its absence. The same study also reported that APEX1 knockout caused mild phenotypic effects, suggesting that perhaps the BER activity detection method was not sensitive enough to rule out a reduced but still sufficient level of BER activity or perhaps that another coordinated BER mechanism might be effective when the subtrate is not an AP site but an oxidized 5mC derivative. Another important study recently showed that NEIL1 and NEIL2 could replace APEX1 function in the TET-catalyzed TDG-dependent active DNA demethylation pathway and their combined knockdown resulted in greater 5fC/5caC accumulation than APEX1 knockdown[216]. Another recent study has suggested that the BER pathway following TDG activity is cell-type dependent, with aformementioned BER components typical of “short-patch repair” participating in demethylation in macrophages while those typical of “long-patch repair” were more commonly involved in active demethylation in neurons[582]. Together, these studies highlight the drawbacks of inferring natural pathways from those which were reconstituted *in vitro* and suggest that future studies could redefine the molecular mechanisms underlying active DNA demethylation.

Concluding remarks

In the work presented in this thesis, I set out to tackle fundamental shortcomings that were broadly characteristic of the field of DNA methylation research in general. I

identified two of these larger issues to be the lack of ability to assess the impact of DNA methylation on gene expression from a causal perspective and a poorly defined understanding of the active DNA demethylation pathway and its constituents. In an endeavour to address these issues, I developed two new technologies by recognizing opportunities to repurpose existing methods towards new applications. In Chapter 2, I repurposed dCas9 binding to interfere with DNMT activity at specific sites and therein optimized and characterized a new system for simple, effective, and specific demethylation of targeted CpGs in dividing cells. I also showed the application of this method at several specific genes, together revealing a variable profile of consequences of demethylation that depend on the proximity of CpGs to transcription start sites and other regulatory elements and on the specific promoter being studied. This represents the most unconfounded method to date to modify DNA methylation levels at specific sites in the genome of live cells and will hopefully serve in the future as a useful method for any researchers in this field to interrogate the causal relationship between DNA demethylation and gene expression across different genomic and physiological contexts. In Chapter 3, I repurposed the EM-seq method by NEB to specifically detect highly oxidized cytosines instead of methylated cytosines, producing a novel, bisulfite-free, simple, efficient, and readily available method for sequencing of these oxidized 5mC derivatives. Armed with this method, I was able to more robustly assess the dynamics of the active DNA demethylation pathway and reported several novel findings, including no transcriptional dependence of TDG activity, a transcriptional activation capacity of TDG, an importance of the full-length TDG protein for demethylation and expression re-activation, a lack of requirement of APEX1 in active DNA demethylation, a striking ubiquitous presence of TDG at active TSS, an interaction between TDG and MBD3/NuRD, and in vivo profiles of CpG oxidation in the mouse cortex. This simplest and most efficient method to discriminate unmethylated cytosines from oxidized cytosines stands to become a useful new tool in the study of active DNA demethylation.

References

1. Sapozhnikov, D.M. & Szyf, M. Increasing Specificity of Targeted DNA Methylation Editing by Non-Enzymatic CRISPR/dCas9-Based Steric Hindrance. *Biomedicines* **11**(2023).
2. Sapozhnikov, D.M. & Szyf, M. Unraveling the functional role of DNA demethylation at specific promoters by targeted steric blockage of DNA methyltransferase with CRISPR/dCas9. *Nat Commun* **12**, 5711 (2021).
3. Sapozhnikov, D.M. & Szyf, M. Enzyme-free targeted DNA demethylation using CRISPR-dCas9-based steric hindrance to identify DNA methylation marks causal to altered gene expression. *Nat Protoc* **17**, 2840-2881 (2022).
4. Ehrlich, M. *et al.* DNA methylation in thermophilic bacteria: N4-methylcytosine, 5-methylcytosine, and N6-methyladenine. *Nucleic Acids Res* **13**, 1399-412 (1985).
5. Luo, G.Z., Blanco, M.A., Greer, E.L., He, C. & Shi, Y. DNA N(6)-methyladenine: a new epigenetic mark in eukaryotes? *Nat Rev Mol Cell Biol* **16**, 705-10 (2015).
6. Boulias, K. & Greer, E.L. Means, mechanisms and consequences of adenine methylation in DNA. *Nat Rev Genet* **23**, 411-428 (2022).
7. Smith, D.W., Crowder, S.W. & Reich, N.O. In vivo specificity of EcoRI DNA methyltransferase. *Nucleic Acids Res* **20**, 6091-6 (1992).
8. Greer, E.L. *et al.* DNA Methylation on N6-Adenine in C. elegans. *Cell* **161**, 868-78 (2015).
9. Zhang, G. *et al.* N6-methyladenine DNA modification in Drosophila. *Cell* **161**, 893-906 (2015).
10. Wu, T.P. *et al.* DNA methylation on N(6)-adenine in mammalian embryonic stem cells. *Nature* **532**, 329-33 (2016).
11. Douvlataniotis, K., Bensberg, M., Lentini, A., Gylemo, B. & Nestor, C.E. No evidence for DNA N (6)-methyladenine in mammals. *Sci Adv* **6**, eaay3335 (2020).
12. Musheev, M.U., Baumgartner, A., Krebs, L. & Niehrs, C. The origin of genomic N(6)-methyl-deoxyadenosine in mammalian cells. *Nat Chem Biol* **16**, 630-634 (2020).
13. Schiffers, S. *et al.* Quantitative LC-MS Provides No Evidence for m(6) dA or m(4) dC in the Genome of Mouse Embryonic Stem Cells and Tissues. *Angew Chem Int Ed Engl* **56**, 11268-11271 (2017).
14. O'Brown, Z.K. *et al.* Sources of artifact in measurements of 6mA and 4mC abundance in eukaryotic genomic DNA. *BMC Genomics* **20**, 445 (2019).
15. Mattei, A.L., Bailly, N. & Meissner, A. DNA methylation: a historical perspective. *Trends Genet* **38**, 676-707 (2022).
16. Niederhuth, C.E. *et al.* Widespread natural variation of DNA methylation within angiosperms. *Genome Biol* **17**, 194 (2016).
17. Ehrlich, M. *et al.* Amount and distribution of 5-methylcytosine in human DNA from different types of tissues of cells. *Nucleic Acids Res* **10**, 2709-21 (1982).
18. Tucker, K.L. Methylated cytosine and the brain: a new base for neuroscience. *Neuron* **30**, 649-52 (2001).
19. Petryk, N., Bultmann, S., Bartke, T. & Defossez, P.A. Staying true to yourself: mechanisms of DNA methylation maintenance in mammals. *Nucleic Acids Res* **49**, 3020-3032 (2021).
20. Rodriguez, F., Yushenova, I.A., DiCorpo, D. & Arkhipova, I.R. Bacterial N4-methylcytosine as an epigenetic mark in eukaryotic DNA. *Nat Commun* **13**, 1072 (2022).

21. Hotchkiss, R.D. The quantitative separation of purines, pyrimidines, and nucleosides by paper chromatography. *J Biol Chem* **175**, 315-32 (1948).
22. Bird, A.P. & Southern, E.M. Use of restriction enzymes to study eukaryotic DNA methylation: I. The methylation pattern in ribosomal DNA from Xenopus laevis. *J Mol Biol* **118**, 27-47 (1978).
23. Waalwijk, C. & Flavell, R.A. MspI, an isoschizomer of hpaII which cleaves both unmethylated and methylated hpaII sites. *Nucleic Acids Res* **5**, 3231-6 (1978).
24. Mohn, F., Weber, M., Schubeler, D. & Roloff, T.C. Methylated DNA immunoprecipitation (MeDIP). *Methods Mol Biol* **507**, 55-64 (2009).
25. Brinkman, A.B. *et al.* Whole-genome DNA methylation profiling using MethylCap-seq. *Methods* **52**, 232-6 (2010).
26. Huang, J., Soupir, A.C. & Wang, L. Cell-free DNA methylome profiling by MBD-seq with ultra-low input. *Epigenetics* **17**, 239-252 (2022).
27. Frommer, M. *et al.* A genomic sequencing protocol that yields a positive display of 5-methylcytosine residues in individual DNA strands. *Proc Natl Acad Sci U S A* **89**, 1827-31 (1992).
28. Shapiro, R., DiFate, V. & Welcher, M. Deamination of cytosine derivatives by bisulfite. Mechanism of the reaction. *J Am Chem Soc* **96**, 906-12 (1974).
29. Grunau, C., Clark, S.J. & Rosenthal, A. Bisulfite genomic sequencing: systematic investigation of critical experimental parameters. *Nucleic Acids Res* **29**, E65-5 (2001).
30. Vaisvila, R. *et al.* Enzymatic methyl sequencing detects DNA methylation at single-base resolution from picograms of DNA. *Genome Res* **31**, 1280-9 (2021).
31. Zhou, J. *et al.* BCREval: a computational method to estimate the bisulfite conversion ratio in WGBS. *BMC Bioinformatics* **21**, 38 (2020).
32. Licchesi, J.D. & Herman, J.G. Methylation-specific PCR. *Methods Mol Biol* **507**, 305-23 (2009).
33. Wojdacz, T.K., Dobrovic, A. & Hansen, L.L. Methylation-sensitive high-resolution melting. *Nat Protoc* **3**, 1903-8 (2008).
34. Li, Y. & Tollefsbol, T.O. DNA methylation detection: bisulfite genomic sequencing analysis. *Methods Mol Biol* **791**, 11-21 (2011).
35. Huang, Z., Bassil, C.F. & Murphy, S.K. Bisulfite sequencing of cloned alleles. *Methods Mol Biol* **1049**, 83-94 (2013).
36. Delaney, C., Garg, S.K. & Yung, R. Analysis of DNA Methylation by Pyrosequencing. *Methods Mol Biol* **1343**, 249-64 (2015).
37. Naik, T., Sharda, M., C, P.L., Virbhadra, K. & Pandit, A. High-quality single amplicon sequencing method for illumina MiSeq platform using pool of 'N' (0-10) spacer-linked target specific primers without PhiX spike-in. *BMC Genomics* **24**, 141 (2023).
38. Olova, N. *et al.* Comparison of whole-genome bisulfite sequencing library preparation strategies identifies sources of biases affecting DNA methylation data. *Genome Biol* **19**, 33 (2018).
39. Gouil, Q. & Keniry, A. Latest techniques to study DNA methylation. *Essays Biochem* **63**, 639-648 (2019).
40. Li, E., Bestor, T.H. & Jaenisch, R. Targeted mutation of the DNA methyltransferase gene results in embryonic lethality. *Cell* **69**, 915-26 (1992).
41. Okano, M., Bell, D.W., Haber, D.A. & Li, E. DNA methyltransferases Dnmt3a and Dnmt3b are essential for de novo methylation and mammalian development. *Cell* **99**, 247-57 (1999).

42. Goyal, R., Reinhardt, R. & Jeltsch, A. Accuracy of DNA methylation pattern preservation by the Dnmt1 methyltransferase. *Nucleic Acids Res* **34**, 1182-8 (2006).
43. Pradhan, S. *et al.* Baculovirus-mediated expression and characterization of the full-length murine DNA methyltransferase. *Nucleic Acids Res* **25**, 4666-73 (1997).
44. Haggerty, C. *et al.* Dnmt1 has de novo activity targeted to transposable elements. *Nat Struct Mol Biol* **28**, 594-603 (2021).
45. Chuang, L.S. *et al.* Human DNA-(cytosine-5) methyltransferase-PCNA complex as a target for p21WAF1. *Science* **277**, 1996-2000 (1997).
46. Iida, T. *et al.* PCNA clamp facilitates action of DNA cytosine methyltransferase 1 on hemimethylated DNA. *Genes Cells* **7**, 997-1007 (2002).
47. Adam, S. *et al.* DNA sequence-dependent activity and base flipping mechanisms of DNMT1 regulate genome-wide DNA methylation. *Nat Commun* **11**, 3723 (2020).
48. Zhou, J. *et al.* Tissue-specific DNA methylation is conserved across human, mouse, and rat, and driven by primary sequence conservation. *BMC Genomics* **18**, 724 (2017).
49. Chen, T., Ueda, Y., Dodge, J.E., Wang, Z. & Li, E. Establishment and maintenance of genomic methylation patterns in mouse embryonic stem cells by Dnmt3a and Dnmt3b. *Mol Cell Biol* **23**, 5594-605 (2003).
50. Feng, J. *et al.* Dnmt1 and Dnmt3a maintain DNA methylation and regulate synaptic function in adult forebrain neurons. *Nat Neurosci* **13**, 423-30 (2010).
51. Liang, G. *et al.* Cooperativity between DNA methyltransferases in the maintenance methylation of repetitive elements. *Mol Cell Biol* **22**, 480-91 (2002).
52. Gujar, H., Weisenberger, D.J. & Liang, G. The Roles of Human DNA Methyltransferases and Their Isoforms in Shaping the Epigenome. *Genes (Basel)* **10**(2019).
53. Lokk, K. *et al.* DNA methylome profiling of human tissues identifies global and tissue-specific methylation patterns. *Genome Biol* **15**, r54 (2014).
54. Leoni, C. *et al.* Dnmt3a restrains mast cell inflammatory responses. *Proc Natl Acad Sci U S A* **114**, E1490-E1499 (2017).
55. Yu, N.K., Baek, S.H. & Kaang, B.K. DNA methylation-mediated control of learning and memory. *Mol Brain* **4**, 5 (2011).
56. Zhang, J., Yang, C., Wu, C., Cui, W. & Wang, L. DNA Methyltransferases in Cancer: Biology, Paradox, Aberrations, and Targeted Therapy. *Cancers (Basel)* **12**(2020).
57. Warhaftig, G. *et al.* Reduction of DNMT3a and RORA in the nucleus accumbens plays a causal role in post-traumatic stress disorder-like behavior: reversal by combinatorial epigenetic therapy. *Mol Psychiatry* **26**, 7481-7497 (2021).
58. Wienholz, B.L. *et al.* DNMT3L modulates significant and distinct flanking sequence preference for DNA methylation by DNMT3A and DNMT3B in vivo. *PLoS Genet* **6**, e1001106 (2010).
59. Suetake, I., Shinozaki, F., Miyagawa, J., Takeshima, H. & Tajima, S. DNMT3L stimulates the DNA methylation activity of Dnmt3a and Dnmt3b through a direct interaction. *J Biol Chem* **279**, 27816-23 (2004).
60. Dukatz, M. *et al.* Complex DNA sequence readout mechanisms of the DNMT3B DNA methyltransferase. *Nucleic Acids Res* **48**, 11495-11509 (2020).
61. Ramsahoye, B.H. *et al.* Non-CpG methylation is prevalent in embryonic stem cells and may be mediated by DNA methyltransferase 3a. *Proc Natl Acad Sci U S A* **97**, 5237-42 (2000).
62. Caldwell, B.A. & Bartolomei, M.S. DNA methylation reprogramming of genomic imprints in the mammalian germline: A TET-centric view. *Andrology* (2022).

63. Kaas, G.A. *et al.* TET1 controls CNS 5-methylcytosine hydroxylation, active DNA demethylation, gene transcription, and memory formation. *Neuron* **79**, 1086-93 (2013).
64. Rudenko, A. *et al.* Tet1 is critical for neuronal activity-regulated gene expression and memory extinction. *Neuron* **79**, 1109-1122 (2013).
65. Miller, C.A. & Sweatt, J.D. Covalent modification of DNA regulates memory formation. *Neuron* **53**, 857-69 (2007).
66. Miller, C.A. *et al.* Cortical DNA methylation maintains remote memory. *Nat Neurosci* **13**, 664-6 (2010).
67. Smith, Z.D. & Meissner, A. The simplest explanation: passive DNA demethylation in PGCs. *EMBO J* **32**, 318-21 (2013).
68. Joshi, K., Liu, S., Breslin, S.J.P. & Zhang, J. Mechanisms that regulate the activities of TET proteins. *Cell Mol Life Sci* **79**, 363 (2022).
69. Rasmussen, K.D. & Helin, K. Role of TET enzymes in DNA methylation, development, and cancer. *Genes Dev* **30**, 733-50 (2016).
70. Bhattacharya, S.K., Ramchandani, S., Cervoni, N. & Szyf, M. A mammalian protein with specific demethylase activity for mCpG DNA. *Nature* **397**, 579-83 (1999).
71. Cao, L. *et al.* Photoelectrochemical biosensor for DNA demethylase detection based on enzymatically induced double-stranded DNA digestion by endonuclease-exonuclease system and Bi(4)O(5)Br(2)-Au/CdS photoactive material. *Talanta* **262**, 124670 (2023).
72. Zhou, Y. *et al.* Enzyme-based electrochemical biosensor for sensitive detection of DNA demethylation and the activity of DNA demethylase. *Anal Chim Acta* **840**, 28-32 (2014).
73. Ng, H.H. *et al.* MBD2 is a transcriptional repressor belonging to the MeCP1 histone deacetylase complex. *Nat Genet* **23**, 58-61 (1999).
74. Roberts, R.J., Vincze, T., Posfai, J. & Macelis, D. REBASE--enzymes and genes for DNA restriction and modification. *Nucleic Acids Res* **35**, D269-70 (2007).
75. Yin, Y. *et al.* Impact of cytosine methylation on DNA binding specificities of human transcription factors. *Science* **356**(2017).
76. Fatemi, M. & Wade, P.A. MBD family proteins: reading the epigenetic code. *J Cell Sci* **119**, 3033-7 (2006).
77. Nan, X., Meehan, R.R. & Bird, A. Dissection of the methyl-CpG binding domain from the chromosomal protein MeCP2. *Nucleic Acids Res* **21**, 4886-92 (1993).
78. Saito, M. & Ishikawa, F. The mCpG-binding domain of human MBD3 does not bind to mCpG but interacts with NuRD/Mi2 components HDAC1 and MTA2. *J Biol Chem* **277**, 35434-9 (2002).
79. Liu, K. *et al.* Structural analyses reveal that MBD3 is a methylated CG binder. *FEBS J* **286**, 3240-3254 (2019).
80. Fournier, A., Sasai, N., Nakao, M. & Defossez, P.A. The role of methyl-binding proteins in chromatin organization and epigenome maintenance. *Brief Funct Genomics* **11**, 251-64 (2012).
81. Klose, R.J. *et al.* DNA binding selectivity of MeCP2 due to a requirement for A/T sequences adjacent to methyl-CpG. *Mol Cell* **19**, 667-78 (2005).
82. Patnaik, D., Esteve, P.O. & Pradhan, S. Targeting the SET and RING-associated (SRA) domain of ubiquitin-like, PHD and ring finger-containing 1 (UHRF1) for anti-cancer drug development. *Oncotarget* **9**, 26243-26258 (2018).
83. Zhou, T. *et al.* Structural basis for hydroxymethylcytosine recognition by the SRA domain of UHRF2. *Mol Cell* **54**, 879-86 (2014).
84. Prokhortchouk, A. *et al.* The p120 catenin partner Kaiso is a DNA methylation-dependent transcriptional repressor. *Genes Dev* **15**, 1613-8 (2001).

85. Filion, G.J. *et al.* A family of human zinc finger proteins that bind methylated DNA and repress transcription. *Mol Cell Biol* **26**, 169-81 (2006).
86. Ruzov, A. *et al.* Kaiso is a genome-wide repressor of transcription that is essential for amphibian development. *Development* **131**, 6185-94 (2004).
87. Bartels, S.J. *et al.* A SILAC-based screen for Methyl-CpG binding proteins identifies RBP-J as a DNA methylation and sequence-specific binding protein. *PLoS One* **6**, e25884 (2011).
88. Hernando-Herraez, I. *et al.* Dynamics of DNA methylation in recent human and great ape evolution. *PLoS Genet* **9**, e1003763 (2013).
89. Scott, C.A. *et al.* Identification of cell type-specific methylation signals in bulk whole genome bisulfite sequencing data. *Genome Biol* **21**, 156 (2020).
90. Lander, E.S. *et al.* Initial sequencing and analysis of the human genome. *Nature* **409**, 860-921 (2001).
91. Xia, J., Han, L. & Zhao, Z. Investigating the relationship of DNA methylation with mutation rate and allele frequency in the human genome. *BMC Genomics* **13 Suppl 8**, S7 (2012).
92. Vinson, C. & Chatterjee, R. CG methylation. *Epigenomics* **4**, 655-63 (2012).
93. Gardiner-Garden, M. & Frommer, M. CpG islands in vertebrate genomes. *J Mol Biol* **196**, 261-82 (1987).
94. Illingworth, R. *et al.* A novel CpG island set identifies tissue-specific methylation at developmental gene loci. *PLoS Biol* **6**, e22 (2008).
95. Carninci, P. *et al.* Genome-wide analysis of mammalian promoter architecture and evolution. *Nat Genet* **38**, 626-35 (2006).
96. Saxonov, S., Berg, P. & Brutlag, D.L. A genome-wide analysis of CpG dinucleotides in the human genome distinguishes two distinct classes of promoters. *Proc Natl Acad Sci U S A* **103**, 1412-7 (2006).
97. Ponger, L., Duret, L. & Mouchiroud, D. Determinants of CpG islands: expression in early embryo and isochore structure. *Genome Res* **11**, 1854-60 (2001).
98. Lim, W.J., Kim, K.H., Kim, J.Y., Jeong, S. & Kim, N. Identification of DNA-Methylated CpG Islands Associated With Gene Silencing in the Adult Body Tissues of the Ogye Chicken Using RNA-Seq and Reduced Representation Bisulfite Sequencing. *Front Genet* **10**, 346 (2019).
99. Shen, L. *et al.* Genome-wide profiling of DNA methylation reveals a class of normally methylated CpG island promoters. *PLoS Genet* **3**, 2023-36 (2007).
100. Ehrlich, M. DNA methylation in cancer: too much, but also too little. *Oncogene* **21**, 5400-13 (2002).
101. Zheng, Y. *et al.* Prediction of genome-wide DNA methylation in repetitive elements. *Nucleic Acids Res* **45**, 8697-8711 (2017).
102. International Human Genome Sequencing, C. Finishing the euchromatic sequence of the human genome. *Nature* **431**, 931-45 (2004).
103. Pappalardo, X.G. & Barra, V. Losing DNA methylation at repetitive elements and breaking bad. *Epigenetics Chromatin* **14**, 25 (2021).
104. Allshire, R.C. & Madhani, H.D. Ten principles of heterochromatin formation and function. *Nat Rev Mol Cell Biol* **19**, 229-244 (2018).
105. Wu, Z. *et al.* DNA methylation modulates HERV-E expression in CD4+ T cells from systemic lupus erythematosus patients. *J Dermatol Sci* **77**, 110-6 (2015).
106. Misiak, B. *et al.* Lower LINE-1 methylation in first-episode schizophrenia patients with the history of childhood trauma. *Epigenomics* **7**, 1275-85 (2015).

107. Natt, D., Johansson, I., Faresjo, T., Ludvigsson, J. & Thorsell, A. High cortisol in 5-year-old children causes loss of DNA methylation in SINE retrotransposons: a possible role for ZNF263 in stress-related diseases. *Clin Epigenetics* **7**, 91 (2015).
108. Criscione, S.W., Zhang, Y., Thompson, W., Sedivy, J.M. & Neretti, N. Transcriptional landscape of repetitive elements in normal and cancer human cells. *BMC Genomics* **15**, 583 (2014).
109. Erichsen, L. *et al.* Genome-wide hypomethylation of LINE-1 and Alu retroelements in cell-free DNA of blood is an epigenetic biomarker of human aging. *Saudi J Biol Sci* **25**, 1220-1226 (2018).
110. Jones, P.A. Functions of DNA methylation: islands, start sites, gene bodies and beyond. *Nat Rev Genet* **13**, 484-92 (2012).
111. Anastasiadi, D., Esteve-Codina, A. & Piferrer, F. Consistent inverse correlation between DNA methylation of the first intron and gene expression across tissues and species. *Epigenetics Chromatin* **11**, 37 (2018).
112. Jeziorska, D.M. *et al.* DNA methylation of intragenic CpG islands depends on their transcriptional activity during differentiation and disease. *Proc Natl Acad Sci U S A* **114**, E7526-E7535 (2017).
113. Maunakea, A.K. *et al.* Conserved role of intragenic DNA methylation in regulating alternative promoters. *Nature* **466**, 253-7 (2010).
114. Lee, S.M. *et al.* Intragenic CpG islands play important roles in bivalent chromatin assembly of developmental genes. *Proc Natl Acad Sci U S A* **114**, E1885-E1894 (2017).
115. Pott, S. & Lieb, J.D. What are super-enhancers? *Nat Genet* **47**, 8-12 (2015).
116. Slieker, R.C. *et al.* Identification and systematic annotation of tissue-specific differentially methylated regions using the Illumina 450k array. *Epigenetics Chromatin* **6**, 26 (2013).
117. Sarda, S., Das, A., Vinson, C. & Hannenhalli, S. Distal CpG islands can serve as alternative promoters to transcribe genes with silenced proximal promoters. *Genome Res* **27**, 553-566 (2017).
118. Waalwijk, C. & Flavell, R.A. DNA methylation at a CCGG sequence in the large intron of the rabbit beta-globin gene: tissue-specific variations. *Nucleic Acids Res* **5**, 4631-4 (1978).
119. Mandel, J.L. & Chambon, P. DNA methylation: organ specific variations in the methylation pattern within and around ovalbumin and other chicken genes. *Nucleic Acids Res* **7**, 2081-103 (1979).
120. Kruczek, I. & Doerfler, W. Expression of the chloramphenicol acetyltransferase gene in mammalian cells under the control of adenovirus type 12 promoters: effect of promoter methylation on gene expression. *Proc Natl Acad Sci U S A* **80**, 7586-90 (1983).
121. Smale, S.T. & Kadonaga, J.T. The RNA polymerase II core promoter. *Annu Rev Biochem* **72**, 449-79 (2003).
122. Slattery, M. *et al.* Absence of a simple code: how transcription factors read the genome. *Trends Biochem Sci* **39**, 381-99 (2014).
123. Kitayner, M. *et al.* Diversity in DNA recognition by p53 revealed by crystal structures with Hoogsteen base pairs. *Nat Struct Mol Biol* **17**, 423-9 (2010).
124. Kuwahara, J., Yonezawa, A., Futamura, M. & Sugiura, Y. Binding of transcription factor Sp1 to GC box DNA revealed by footprinting analysis: different contact of three zinc fingers and sequence recognition mode. *Biochemistry* **32**, 5994-6001 (1993).
125. Chen, H. & Pugh, B.F. What do Transcription Factors Interact With? *J Mol Biol* **433**, 166883 (2021).

126. Dantas Machado, A.C. *et al.* Evolving insights on how cytosine methylation affects protein-DNA binding. *Brief Funct Genomics* **14**, 61-73 (2015).
127. Hellman, L.M. & Fried, M.G. Electrophoretic mobility shift assay (EMSA) for detecting protein-nucleic acid interactions. *Nat Protoc* **2**, 1849-61 (2007).
128. Li, Y., Xiao, D., Yang, S. & Zhang, L. Promoter methylation represses AT2R gene and increases brain hypoxic-ischemic injury in neonatal rats. *Neurobiol Dis* **60**, 32-8 (2013).
129. Weaver, I.C., Diorio, J., Seckl, J.R., Szyf, M. & Meaney, M.J. Early environmental regulation of hippocampal glucocorticoid receptor gene expression: characterization of intracellular mediators and potential genomic target sites. *Ann N Y Acad Sci* **1024**, 182-212 (2004).
130. Gallinari, P., Di Marco, S., Jones, P., Pallaoro, M. & Steinkuhler, C. HDACs, histone deacetylation and gene transcription: from molecular biology to cancer therapeutics. *Cell Res* **17**, 195-211 (2007).
131. Lee, D.Y., Hayes, J.J., Pruss, D. & Wolffe, A.P. A positive role for histone acetylation in transcription factor access to nucleosomal DNA. *Cell* **72**, 73-84 (1993).
132. Kaluscha, S. *et al.* Evidence that direct inhibition of transcription factor binding is the prevailing mode of gene and repeat repression by DNA methylation. *Nat Genet* **54**, 1895-1906 (2022).
133. Kuroda, A. *et al.* Insulin gene expression is regulated by DNA methylation. *PLoS One* **4**, e6953 (2009).
134. Takizawa, T. *et al.* DNA methylation is a critical cell-intrinsic determinant of astrocyte differentiation in the fetal brain. *Dev Cell* **1**, 749-58 (2001).
135. Yagi, S. *et al.* DNA methylation profile of tissue-dependent and differentially methylated regions (T-DMRs) in mouse promoter regions demonstrating tissue-specific gene expression. *Genome Res* **18**, 1969-78 (2008).
136. Steele, R.J., Thompson, A.M., Hall, P.A. & Lane, D.P. The p53 tumour suppressor gene. *Br J Surg* **85**, 1460-7 (1998).
137. Yeo, C.J. Tumor suppressor genes: a short review. *Surgery* **125**, 363-6 (1999).
138. Herman, J.G. *et al.* Silencing of the VHL tumor-suppressor gene by DNA methylation in renal carcinoma. *Proc Natl Acad Sci U S A* **91**, 9700-4 (1994).
139. Costello, J.F., Berger, M.S., Huang, H.S. & Cavenee, W.K. Silencing of p16/CDKN2 expression in human gliomas by methylation and chromatin condensation. *Cancer Res* **56**, 2405-10 (1996).
140. Chiang, J.W., Karlan, B.Y., Cass, L. & Baldwin, R.L. BRCA1 promoter methylation predicts adverse ovarian cancer prognosis. *Gynecol Oncol* **101**, 403-10 (2006).
141. Alvarez-Nunez, F. *et al.* PTEN promoter methylation in sporadic thyroid carcinomas. *Thyroid* **16**, 17-23 (2006).
142. Li, X. *et al.* MLH1 promoter methylation frequency in colorectal cancer patients and related clinicopathological and molecular features. *PLoS One* **8**, e59064 (2013).
143. Agirre, X. *et al.* Epigenetic silencing of the tumor suppressor microRNA Hsa-miR-124a regulates CDK6 expression and confers a poor prognosis in acute lymphoblastic leukemia. *Cancer Res* **69**, 4443-53 (2009).
144. Ferreira, H.J., Heyn, H., Moutinho, C. & Esteller, M. CpG island hypermethylation-associated silencing of small nucleolar RNAs in human cancer. *RNA Biol* **9**, 881-90 (2012).
145. Montero, A.J. *et al.* Epigenetic inactivation of EGFR by CpG island hypermethylation in cancer. *Cancer Biol Ther* **5**, 1494-501 (2006).

146. Feinberg, A.P. & Vogelstein, B. Hypomethylation of ras oncogenes in primary human cancers. *Biochem Biophys Res Commun* **111**, 47-54 (1983).
147. de Souza, C.R. *et al.* MYC deregulation in gastric cancer and its clinicopathological implications. *PLoS One* **8**, e64420 (2013).
148. Couronne, L., Bastard, C. & Bernard, O.A. TET2 and DNMT3A mutations in human T-cell lymphoma. *N Engl J Med* **366**, 95-6 (2012).
149. Yang, B.T. *et al.* Insulin promoter DNA methylation correlates negatively with insulin gene expression and positively with HbA(1c) levels in human pancreatic islets. *Diabetologia* **54**, 360-7 (2011).
150. Sailani, M.R. *et al.* Lifelong physical activity is associated with promoter hypomethylation of genes involved in metabolism, myogenesis, contractile properties and oxidative stress resistance in aged human skeletal muscle. *Sci Rep* **9**, 3272 (2019).
151. Restrepo, A. *et al.* Epigenetic regulation of glial fibrillary acidic protein by DNA methylation in human malignant gliomas. *Neuro Oncol* **13**, 42-50 (2011).
152. Ahmed, O.G. *et al.* Folic acid ameliorates neonatal isolation-induced autistic like behaviors in rats: epigenetic modifications of BDNF and GFAP promotors. *Appl Physiol Nutr Metab* **46**, 964-975 (2021).
153. Good, M. *et al.* Global hypermethylation of intestinal epithelial cells is a hallmark feature of neonatal surgical necrotizing enterocolitis. *Clin Epigenetics* **12**, 190 (2020).
154. Yamaguchi, K. *et al.* Epigenetic determinants of ovarian clear cell carcinoma biology. *Int J Cancer* **135**, 585-97 (2014).
155. McGowan, P.O. *et al.* Epigenetic regulation of the glucocorticoid receptor in human brain associates with childhood abuse. *Nat Neurosci* **12**, 342-8 (2009).
156. Kim, A.Y. *et al.* Obesity-induced DNA hypermethylation of the adiponectin gene mediates insulin resistance. *Nat Commun* **6**, 7585 (2015).
157. Jowaed, A., Schmitt, I., Kaut, O. & Wullner, U. Methylation regulates alpha-synuclein expression and is decreased in Parkinson's disease patients' brains. *J Neurosci* **30**, 6355-9 (2010).
158. Ball, M.P. *et al.* Targeted and genome-scale strategies reveal gene-body methylation signatures in human cells. *Nat Biotechnol* **27**, 361-8 (2009).
159. Wolf, S.F., Jolly, D.J., Lunnen, K.D., Friedmann, T. & Migeon, B.R. Methylation of the hypoxanthine phosphoribosyltransferase locus on the human X chromosome: implications for X-chromosome inactivation. *Proc Natl Acad Sci U S A* **81**, 2806-10 (1984).
160. Hellman, A. & Chess, A. Gene body-specific methylation on the active X chromosome. *Science* **315**, 1141-3 (2007).
161. Laurent, L. *et al.* Dynamic changes in the human methylome during differentiation. *Genome Res* **20**, 320-31 (2010).
162. Neri, F. *et al.* Intragenic DNA methylation prevents spurious transcription initiation. *Nature* **543**, 72-77 (2017).
163. Kimura, K. *et al.* Diversification of transcriptional modulation: large-scale identification and characterization of putative alternative promoters of human genes. *Genome Res* **16**, 55-65 (2006).
164. Cinghu, S. *et al.* Intragenic Enhancers Attenuate Host Gene Expression. *Mol Cell* **68**, 104-117 e6 (2017).
165. Birnbaum, R.Y. *et al.* Coding exons function as tissue-specific enhancers of nearby genes. *Genome Res* **22**, 1059-68 (2012).

166. Borsari, B. *et al.* Enhancers with tissue-specific activity are enriched in intronic regions. *Genome Res* **31**, 1325-1336 (2021).
167. Coppola, C.J., R, C.R. & Mendenhall, E.M. Identification and function of enhancers in the human genome. *Hum Mol Genet* **25**, R190-R197 (2016).
168. Ong, C.T. & Corces, V.G. Enhancer function: new insights into the regulation of tissue-specific gene expression. *Nat Rev Genet* **12**, 283-93 (2011).
169. Sen, P. *et al.* Spurious intragenic transcription is a feature of mammalian cellular senescence and tissue aging. *Nat Aging* **3**, 402-417 (2023).
170. Wang, Q. *et al.* Gene body methylation in cancer: molecular mechanisms and clinical applications. *Clin Epigenetics* **14**, 154 (2022).
171. de la Mata, M. *et al.* A slow RNA polymerase II affects alternative splicing in vivo. *Mol Cell* **12**, 525-32 (2003).
172. Chathoth, K.T., Barrass, J.D., Webb, S. & Beggs, J.D. A splicing-dependent transcriptional checkpoint associated with prespliceosome formation. *Mol Cell* **53**, 779-90 (2014).
173. Kornblihtt, A.R., de la Mata, M., Fededa, J.P., Munoz, M.J. & Nogues, G. Multiple links between transcription and splicing. *RNA* **10**, 1489-98 (2004).
174. Veloso, A. *et al.* Rate of elongation by RNA polymerase II is associated with specific gene features and epigenetic modifications. *Genome Res* **24**, 896-905 (2014).
175. Lev Maor, G., Yearim, A. & Ast, G. The alternative role of DNA methylation in splicing regulation. *Trends Genet* **31**, 274-80 (2015).
176. Shukla, S. *et al.* CTCF-promoted RNA polymerase II pausing links DNA methylation to splicing. *Nature* **479**, 74-9 (2011).
177. Maunakea, A.K., Chepelev, I., Cui, K. & Zhao, K. Intragenic DNA methylation modulates alternative splicing by recruiting MeCP2 to promote exon recognition. *Cell Res* **23**, 1256-69 (2013).
178. Young, J.I. *et al.* Regulation of RNA splicing by the methylation-dependent transcriptional repressor methyl-CpG binding protein 2. *Proc Natl Acad Sci U S A* **102**, 17551-8 (2005).
179. Yearim, A. *et al.* HP1 is involved in regulating the global impact of DNA methylation on alternative splicing. *Cell Rep* **10**, 1122-34 (2015).
180. Williams, K., Christensen, J. & Helin, K. DNA methylation: TET proteins-guardians of CpG islands? *EMBO Rep* **13**, 28-35 (2011).
181. Teissandier, A. & Bourc'his, D. Gene body DNA methylation conspires with H3K36me3 to preclude aberrant transcription. *EMBO J* **36**, 1471-1473 (2017).
182. Miller, J.L. & Grant, P.A. The role of DNA methylation and histone modifications in transcriptional regulation in humans. *Subcell Biochem* **61**, 289-317 (2013).
183. Beacon, T.H. *et al.* The dynamic broad epigenetic (H3K4me3, H3K27ac) domain as a mark of essential genes. *Clin Epigenetics* **13**, 138 (2021).
184. Meissner, A. *et al.* Genome-scale DNA methylation maps of pluripotent and differentiated cells. *Nature* **454**, 766-70 (2008).
185. Fuks, F. *et al.* The methyl-CpG-binding protein MeCP2 links DNA methylation to histone methylation. *J Biol Chem* **278**, 4035-40 (2003).
186. Nan, X. *et al.* Transcriptional repression by the methyl-CpG-binding protein MeCP2 involves a histone deacetylase complex. *Nature* **393**, 386-9 (1998).
187. Nott, A. *et al.* Histone deacetylase 3 associates with MeCP2 to regulate FOXO and social behavior. *Nat Neurosci* **19**, 1497-1505 (2016).

188. Jones, P.L. *et al.* Methylated DNA and MeCP2 recruit histone deacetylase to repress transcription. *Nat Genet* **19**, 187-91 (1998).
189. Zhang, Y. *et al.* Analysis of the NuRD subunits reveals a histone deacetylase core complex and a connection with DNA methylation. *Genes Dev* **13**, 1924-35 (1999).
190. Fujita, N. *et al.* Methyl-CpG binding domain 1 (MBD1) interacts with the Suv39h1-HP1 heterochromatic complex for DNA methylation-based transcriptional repression. *J Biol Chem* **278**, 24132-8 (2003).
191. Otani, J. *et al.* Structural basis for recognition of H3K4 methylation status by the DNA methyltransferase 3A ATRX-DNMT3-DNMT3L domain. *EMBO Rep* **10**, 1235-41 (2009).
192. Rose, N.R. & Klose, R.J. Understanding the relationship between DNA methylation and histone lysine methylation. *Biochim Biophys Acta* **1839**, 1362-72 (2014).
193. Agarwal, N. *et al.* MeCP2 interacts with HP1 and modulates its heterochromatin association during myogenic differentiation. *Nucleic Acids Res* **35**, 5402-8 (2007).
194. Smallwood, A., Esteve, P.O., Pradhan, S. & Carey, M. Functional cooperation between HP1 and DNMT1 mediates gene silencing. *Genes Dev* **21**, 1169-78 (2007).
195. Fuks, F., Hurd, P.J., Deplus, R. & Kouzarides, T. The DNA methyltransferases associate with HP1 and the SUV39H1 histone methyltransferase. *Nucleic Acids Res* **31**, 2305-12 (2003).
196. Vire, E. *et al.* The Polycomb group protein EZH2 directly controls DNA methylation. *Nature* **439**, 871-4 (2006).
197. Li, Y., Chen, X. & Lu, C. The interplay between DNA and histone methylation: molecular mechanisms and disease implications. *EMBO Rep* **22**, e51803 (2021).
198. Stirzaker, C., Song, J.Z., Davidson, B. & Clark, S.J. Transcriptional gene silencing promotes DNA hypermethylation through a sequential change in chromatin modifications in cancer cells. *Cancer Res* **64**, 3871-7 (2004).
199. Padjen, K., Ratnam, S. & Storb, U. DNA methylation precedes chromatin modifications under the influence of the strain-specific modifier Ssm1. *Mol Cell Biol* **25**, 4782-91 (2005).
200. Stewart, K.R. *et al.* Dynamic changes in histone modifications precede de novo DNA methylation in oocytes. *Genes Dev* **29**, 2449-62 (2015).
201. Mochizuki, K. *et al.* Repression of germline genes by PRC1.6 and SETDB1 in the early embryo precedes DNA methylation-mediated silencing. *Nat Commun* **12**, 7020 (2021).
202. Mutskov, V. & Felsenfeld, G. Silencing of transgene transcription precedes methylation of promoter DNA and histone H3 lysine 9. *EMBO J* **23**, 138-49 (2004).
203. Pacis, A. *et al.* Gene activation precedes DNA demethylation in response to infection in human dendritic cells. *Proc Natl Acad Sci U S A* **116**, 6938-6943 (2019).
204. Chakraborty, R. *et al.* Histone Acetyltransferases p300 and CBP Coordinate Distinct Chromatin Remodeling Programs in Vascular Smooth Muscle Plasticity. *Circulation* **145**, 1720-1737 (2022).
205. Lu, H.G. *et al.* TET1 partially mediates HDAC inhibitor-induced suppression of breast cancer invasion. *Mol Med Rep* **10**, 2595-600 (2014).
206. Chen, Q., Chen, Y., Bian, C., Fujiki, R. & Yu, X. TET2 promotes histone O-GlcNAcylation during gene transcription. *Nature* **493**, 561-4 (2013).
207. Tahiliani, M. *et al.* Conversion of 5-methylcytosine to 5-hydroxymethylcytosine in mammalian DNA by MLL partner TET1. *Science* **324**, 930-5 (2009).
208. Ito, S. *et al.* Role of Tet proteins in 5mC to 5hmC conversion, ES-cell self-renewal and inner cell mass specification. *Nature* **466**, 1129-33 (2010).

209. He, Y.F. *et al.* Tet-mediated formation of 5-carboxylcytosine and its excision by TDG in mammalian DNA. *Science* **333**, 1303-7 (2011).
210. Hu, L. *et al.* Crystal structure of TET2-DNA complex: insight into TET-mediated 5mC oxidation. *Cell* **155**, 1545-55 (2013).
211. Crawford, D.J. *et al.* Tet2 Catalyzes Stepwise 5-Methylcytosine Oxidation by an Iterative and de novo Mechanism. *J Am Chem Soc* **138**, 730-3 (2016).
212. Weber, A.R. *et al.* Biochemical reconstitution of TET1-TDG-BER-dependent active DNA demethylation reveals a highly coordinated mechanism. *Nat Commun* **7**, 10806 (2016).
213. Maiti, A. & Drohat, A.C. Thymine DNA glycosylase can rapidly excise 5-formylcytosine and 5-carboxylcytosine: potential implications for active demethylation of CpG sites. *J Biol Chem* **286**, 35334-35338 (2011).
214. Hashimoto, H., Hong, S., Bhagwat, A.S., Zhang, X. & Cheng, X. Excision of 5-hydroxymethyluracil and 5-carboxylcytosine by the thymine DNA glycosylase domain: its structural basis and implications for active DNA demethylation. *Nucleic Acids Res* **40**, 10203-14 (2012).
215. Krokan, H.E. & Bjoras, M. Base excision repair. *Cold Spring Harb Perspect Biol* **5**, a012583 (2013).
216. Schomacher, L. *et al.* Neil DNA glycosylases promote substrate turnover by Tdg during DNA demethylation. *Nat Struct Mol Biol* **23**, 116-124 (2016).
217. Whitaker, A.M. & Freudenthal, B.D. APE1: A skilled nucleic acid surgeon. *DNA Repair (Amst)* **71**, 93-100 (2018).
218. Szyf, M. The elusive role of 5'-hydroxymethylcytosine. *Epigenomics* **8**, 1539-1551 (2016).
219. Wu, X. & Zhang, Y. TET-mediated active DNA demethylation: mechanism, function and beyond. *Nat Rev Genet* **18**, 517-534 (2017).
220. Lee, H.J., Hore, T.A. & Reik, W. Reprogramming the methylome: erasing memory and creating diversity. *Cell Stem Cell* **14**, 710-9 (2014).
221. Guo, F. *et al.* Active and passive demethylation of male and female pronuclear DNA in the mammalian zygote. *Cell Stem Cell* **15**, 447-459 (2014).
222. Wossidlo, M. *et al.* 5-Hydroxymethylcytosine in the mammalian zygote is linked with epigenetic reprogramming. *Nat Commun* **2**, 241 (2011).
223. Iqbal, K., Jin, S.G., Pfeifer, G.P. & Szabo, P.E. Reprogramming of the paternal genome upon fertilization involves genome-wide oxidation of 5-methylcytosine. *Proc Natl Acad Sci U S A* **108**, 3642-7 (2011).
224. Inoue, A., Shen, L., Dai, Q., He, C. & Zhang, Y. Generation and replication-dependent dilution of 5fC and 5caC during mouse preimplantation development. *Cell Res* **21**, 1670-6 (2011).
225. Inoue, A., Shen, L., Matoba, S. & Zhang, Y. Haploinsufficiency, but not defective paternal 5mC oxidation, accounts for the developmental defects of maternal Tet3 knockouts. *Cell Rep* **10**, 463-70 (2015).
226. Tang, F. *et al.* Deterministic and stochastic allele specific gene expression in single mouse blastomeres. *PLoS One* **6**, e21208 (2011).
227. Ishida, M. & Moore, G.E. The role of imprinted genes in humans. *Mol Aspects Med* **34**, 826-40 (2013).
228. Kempermann, G. *et al.* Human Adult Neurogenesis: Evidence and Remaining Questions. *Cell Stem Cell* **23**, 25-30 (2018).
229. Martinowich, K. *et al.* DNA methylation-related chromatin remodeling in activity-dependent BDNF gene regulation. *Science* **302**, 890-3 (2003).

230. Ma, D.K. *et al.* Neuronal activity-induced Gadd45b promotes epigenetic DNA demethylation and adult neurogenesis. *Science* **323**, 1074-7 (2009).
231. Guo, J.U. *et al.* Neuronal activity modifies the DNA methylation landscape in the adult brain. *Nat Neurosci* **14**, 1345-51 (2011).
232. 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**, 423-34 (2011).
233. Ito, S. *et al.* Tet proteins can convert 5-methylcytosine to 5-formylcytosine and 5-carboxylcytosine. *Science* **333**, 1300-3 (2011).
234. Sagarkar, S. *et al.* TET1-induced DNA demethylation in dentate gyrus is important for reward conditioning and reinforcement. *Mol Neurobiol* **59**, 5426-5442 (2022).
235. Li, X. *et al.* Neocortical Tet3-mediated accumulation of 5-hydroxymethylcytosine promotes rapid behavioral adaptation. *Proc Natl Acad Sci U S A* **111**, 7120-5 (2014).
236. Yu, H. *et al.* Tet3 regulates synaptic transmission and homeostatic plasticity via DNA oxidation and repair. *Nat Neurosci* **18**, 836-43 (2015).
237. Chrysanthou, S. *et al.* The DNA dioxygenase Tet1 regulates H3K27 modification and embryonic stem cell biology independent of its catalytic activity. *Nucleic Acids Res* **50**, 3169-3189 (2022).
238. Zhang, R.R. *et al.* Tet1 regulates adult hippocampal neurogenesis and cognition. *Cell Stem Cell* **13**, 237-45 (2013).
239. Bordin, D.L., Lirussi, L. & Nilsen, H. Cellular response to endogenous DNA damage: DNA base modifications in gene expression regulation. *DNA Repair (Amst)* **99**, 103051 (2021).
240. Booth, M.J. *et al.* Oxidative bisulfite sequencing of 5-methylcytosine and 5-hydroxymethylcytosine. *Nat Protoc* **8**, 1841-51 (2013).
241. Bachman, M. *et al.* 5-Hydroxymethylcytosine is a predominantly stable DNA modification. *Nat Chem* **6**, 1049-55 (2014).
242. Bachman, M. *et al.* 5-Formylcytosine can be a stable DNA modification in mammals. *Nat Chem Biol* **11**, 555-7 (2015).
243. Wang, T. *et al.* Direct enzymatic sequencing of 5-methylcytosine at single-base resolution. *Nat Chem Biol* (2023).
244. Neri, F. *et al.* Single-Base Resolution Analysis of 5-Formyl and 5-Carboxyl Cytosine Reveals Promoter DNA Methylation Dynamics. *Cell Rep* **10**, 674-683 (2015).
245. Lu, X. *et al.* Chemical modification-assisted bisulfite sequencing (CAB-Seq) for 5-carboxylcytosine detection in DNA. *J Am Chem Soc* **135**, 9315-7 (2013).
246. Lu, X. *et al.* Base-resolution maps of 5-formylcytosine and 5-carboxylcytosine reveal genome-wide DNA demethylation dynamics. *Cell Res* **25**, 386-9 (2015).
247. Licyte, J. *et al.* A Bisulfite-free Approach for Base-Resolution Analysis of Genomic 5-Carboxylcytosine. *Cell Rep* **32**, 108155 (2020).
248. He, B. *et al.* Tissue-specific 5-hydroxymethylcytosine landscape of the human genome. *Nat Commun* **12**, 4249 (2021).
249. Borkowska, J. *et al.* Alterations in 5hmC level and genomic distribution in aging-related epigenetic drift in human adipose stem cells. *Epigenomics* **12**, 423-437 (2020).
250. Pastor, W.A. *et al.* Genome-wide mapping of 5-hydroxymethylcytosine in embryonic stem cells. *Nature* **473**, 394-7 (2011).
251. Sun, Z. *et al.* A sensitive approach to map genome-wide 5-hydroxymethylcytosine and 5-formylcytosine at single-base resolution. *Mol Cell* **57**, 750-761 (2015).
252. Song, C.X. *et al.* Selective chemical labeling reveals the genome-wide distribution of 5-hydroxymethylcytosine. *Nat Biotechnol* **29**, 68-72 (2011).

253. Wu, H. *et al.* Genome-wide analysis of 5-hydroxymethylcytosine distribution reveals its dual function in transcriptional regulation in mouse embryonic stem cells. *Genes Dev* **25**, 679-84 (2011).
254. Rodriguez-Aguilera, J.R. *et al.* Genome-wide 5-hydroxymethylcytosine (5hmC) emerges at early stage of in vitro differentiation of a putative hepatocyte progenitor. *Sci Rep* **10**, 7822 (2020).
255. Jin, S.G., Wu, X., Li, A.X. & Pfeifer, G.P. Genomic mapping of 5-hydroxymethylcytosine in the human brain. *Nucleic Acids Res* **39**, 5015-24 (2011).
256. Booth, M.J., Marsico, G., Bachman, M., Beraldi, D. & Balasubramanian, S. Quantitative sequencing of 5-formylcytosine in DNA at single-base resolution. *Nat Chem* **6**, 435-40 (2014).
257. Song, C.X. *et al.* Genome-wide profiling of 5-formylcytosine reveals its roles in epigenetic priming. *Cell* **153**, 678-91 (2013).
258. Xia, B. *et al.* Bisulfite-free, base-resolution analysis of 5-formylcytosine at the genome scale. *Nat Methods* **12**, 1047-50 (2015).
259. Zhu, C. *et al.* Single-Cell 5-Formylcytosine Landscapes of Mammalian Early Embryos and ESCs at Single-Base Resolution. *Cell Stem Cell* **20**, 720-731 e5 (2017).
260. Wheldon, L.M. *et al.* Transient accumulation of 5-carboxylcytosine indicates involvement of active demethylation in lineage specification of neural stem cells. *Cell Rep* **7**, 1353-1361 (2014).
261. Nanan, K.K. *et al.* TET-Catalyzed 5-Carboxylcytosine Promotes CTCF Binding to Suboptimal Sequences Genome-wide. *iScience* **19**, 326-339 (2019).
262. Iurlaro, M. *et al.* A screen for hydroxymethylcytosine and formylcytosine binding proteins suggests functions in transcription and chromatin regulation. *Genome Biol* **14**, R119 (2013).
263. Yildirim, O. *et al.* Mbd3/NURD complex regulates expression of 5-hydroxymethylcytosine marked genes in embryonic stem cells. *Cell* **147**, 1498-510 (2011).
264. Cusack, M. *et al.* Distinct contributions of DNA methylation and histone acetylation to the genomic occupancy of transcription factors. *Genome Res* **30**, 1393-1406 (2020).
265. Da, M.X., Zhang, Y.B., Yao, J.B. & Duan, Y.X. DNA methylation regulates expression of VEGF-C, and S-adenosylmethionine is effective for VEGF-C methylation and for inhibiting cancer growth. *Braz J Med Biol Res* **47**, 1021-8 (2014).
266. Yang, H. *et al.* Role of promoter methylation in increased methionine adenosyltransferase 2A expression in human liver cancer. *Am J Physiol Gastrointest Liver Physiol* **280**, G184-90 (2001).
267. Petrossian, T.C. & Clarke, S.G. Uncovering the human methyltransferasome. *Mol Cell Proteomics* **10**, M110 000976 (2011).
268. Ouyang, Y., Wu, Q., Li, J., Sun, S. & Sun, S. S-adenosylmethionine: A metabolite critical to the regulation of autophagy. *Cell Prolif* **53**, e12891 (2020).
269. Taylor, S.M. & Jones, P.A. Mechanism of action of eukaryotic DNA methyltransferase. Use of 5-azacytosine-containing DNA. *J Mol Biol* **162**, 679-92 (1982).
270. Santi, D.V., Norment, A. & Garrett, C.E. Covalent bond formation between a DNA-cytosine methyltransferase and DNA containing 5-azacytosine. *Proc Natl Acad Sci U S A* **81**, 6993-7 (1984).
271. Gabbara, S. & Bhagwat, A.S. The mechanism of inhibition of DNA (cytosine-5-)-methyltransferases by 5-azacytosine is likely to involve methyl transfer to the inhibitor. *Biochem J* **307 (Pt 1)**, 87-92 (1995).

272. Creusot, F., Acs, G. & Christman, J.K. Inhibition of DNA methyltransferase and induction of Friend erythroleukemia cell differentiation by 5-azacytidine and 5-aza-2'-deoxycytidine. *J Biol Chem* **257**, 2041-8 (1982).
273. Christman, J.K., Mendelsohn, N., Herzog, D. & Schneiderman, N. Effect of 5-azacytidine on differentiation and DNA methylation in human promyelocytic leukemia cells (HL-60). *Cancer Res* **43**, 763-9 (1983).
274. Lu, S. & Davies, P.J. Regulation of the expression of the tissue transglutaminase gene by DNA methylation. *Proc Natl Acad Sci U S A* **94**, 4692-7 (1997).
275. Chan, Y. *et al.* The cell-specific expression of endothelial nitric-oxide synthase: a role for DNA methylation. *J Biol Chem* **279**, 35087-100 (2004).
276. Kuroda, M. *et al.* DNA Methylation Suppresses Leptin Gene in 3T3-L1 Adipocytes. *PLoS One* **11**, e0160532 (2016).
277. Matt, S.M., Lawson, M.A. & Johnson, R.W. Aging and peripheral lipopolysaccharide can modulate epigenetic regulators and decrease IL-1beta promoter DNA methylation in microglia. *Neurobiol Aging* **47**, 1-9 (2016).
278. Karouzakis, E. *et al.* DNA methylation regulates the expression of CXCL12 in rheumatoid arthritis synovial fibroblasts. *Genes Immun* **12**, 643-52 (2011).
279. Kusui, C. *et al.* DNA methylation of the human oxytocin receptor gene promoter regulates tissue-specific gene suppression. *Biochem Biophys Res Commun* **289**, 681-6 (2001).
280. Juttermann, R., Li, E. & Jaenisch, R. Toxicity of 5-aza-2'-deoxycytidine to mammalian cells is mediated primarily by covalent trapping of DNA methyltransferase rather than DNA demethylation. *Proc Natl Acad Sci U S A* **91**, 11797-801 (1994).
281. Seelan, R.S., Mukhopadhyay, P., Pisano, M.M. & Greene, R.M. Effects of 5-Aza-2'-deoxycytidine (decitabine) on gene expression. *Drug Metab Rev* **50**, 193-207 (2018).
282. Kondo, Y., Shen, L. & Issa, J.P. Critical role of histone methylation in tumor suppressor gene silencing in colorectal cancer. *Mol Cell Biol* **23**, 206-15 (2003).
283. Azevedo Portilho, N. *et al.* The DNMT1 inhibitor GSK-3484862 mediates global demethylation in murine embryonic stem cells. *Epigenetics Chromatin* **14**, 56 (2021).
284. Hu, B. *et al.* Therapeutic siRNA: state of the art. *Signal Transduct Target Ther* **5**, 101 (2020).
285. Robert, M.F. *et al.* DNMT1 is required to maintain CpG methylation and aberrant gene silencing in human cancer cells. *Nat Genet* **33**, 61-5 (2003).
286. Dhawan, S., Georgia, S., Tschen, S.I., Fan, G. & Bhushan, A. Pancreatic beta cell identity is maintained by DNA methylation-mediated repression of Arx. *Dev Cell* **20**, 419-29 (2011).
287. Lai, Q. *et al.* The loss-of-function of DNA methyltransferase 1 by siRNA impairs the growth of non-small cell lung cancer with alleviated side effects via reactivation of RASSF1A and APC in vitro and vivo. *Oncotarget* **8**, 59301-59311 (2017).
288. Ning, X. *et al.* DNMT1 and EZH2 mediated methylation silences the microRNA-200b/a/429 gene and promotes tumor progression. *Cancer Lett* **359**, 198-205 (2015).
289. Xu, D. *et al.* DNMT1 mediated promoter methylation of GNAO1 in hepatoma carcinoma cells. *Gene* **665**, 67-73 (2018).
290. Ferguson, A.T., Lapidus, R.G. & Davidson, N.E. Demethylation of the progesterone receptor CpG island is not required for progesterone receptor gene expression. *Oncogene* **17**, 577-83 (1998).
291. Cheishvili, D. *et al.* DNA methylation controls unmethylated transcription start sites in the genome in trans. *Epigenomics* **9**, 611-633 (2017).

292. Murakami, J. *et al.* Effects of demethylating agent 5-aza-2(')-deoxycytidine and histone deacetylase inhibitor FR901228 on maspin gene expression in oral cancer cell lines. *Oral Oncol* **40**, 597-603 (2004).
293. de Silva, S. *et al.* Promoter methylation regulates SAMHD1 gene expression in human CD4+ T cells. *J Biol Chem* **288**, 9284-92 (2013).
294. Policicchio, S. *et al.* Genome-wide DNA methylation meta-analysis in the brains of suicide completers. *Transl Psychiatry* **10**, 69 (2020).
295. Polansky, J.K. *et al.* DNA methylation controls Foxp3 gene expression. *Eur J Immunol* **38**, 1654-63 (2008).
296. Chandra, A., Senapati, S., Roy, S., Chatterjee, G. & Chatterjee, R. Epigenome-wide DNA methylation regulates cardinal pathological features of psoriasis. *Clin Epigenetics* **10**, 108 (2018).
297. Ghoshal, K. *et al.* Role of human ribosomal RNA (rRNA) promoter methylation and of methyl-CpG-binding protein MBD2 in the suppression of rRNA gene expression. *J Biol Chem* **279**, 6783-93 (2004).
298. Yang, X. *et al.* DNA Methylation Biphasically Regulates 3T3-L1 Preadipocyte Differentiation. *Mol Endocrinol* **30**, 677-87 (2016).
299. Klug, M. & Rehli, M. Functional analysis of promoter CpG methylation using a CpG-free luciferase reporter vector. *Epigenetics* **1**, 127-30 (2006).
300. Lu, Q., Ray, D., Gutsch, D. & Richardson, B. Effect of DNA methylation and chromatin structure on ITGAL expression. *Blood* **99**, 4503-8 (2002).
301. Curradi, M., Izzo, A., Badaracco, G. & Landsberger, N. Molecular mechanisms of gene silencing mediated by DNA methylation. *Mol Cell Biol* **22**, 3157-73 (2002).
302. Han, W., Shi, M. & Spivack, S.D. Site-specific methylated reporter constructs for functional analysis of DNA methylation. *Epigenetics* **8**, 1176-87 (2013).
303. Waryah, C.B., Moses, C., Arooj, M. & Blancafort, P. Zinc Fingers, TALEs, and CRISPR Systems: A Comparison of Tools for Epigenome Editing. *Methods Mol Biol* **1767**, 19-63 (2018).
304. Qi, L.S. *et al.* Repurposing CRISPR as an RNA-guided platform for sequence-specific control of gene expression. *Cell* **152**, 1173-83 (2013).
305. Konermann, S. *et al.* Genome-scale transcriptional activation by an engineered CRISPR-Cas9 complex. *Nature* **517**, 583-8 (2015).
306. Shalem, O. *et al.* Genome-Scale CRISPR-Cas9 Knockout Screening in Human Cells. *Science* **343**, 84-87 (2014).
307. Wegmann, S. *et al.* Persistent repression of tau in the brain using engineered zinc finger protein transcription factors. *Sci Adv* **7**(2021).
308. Li, H. *et al.* In vivo genome editing restores haemostasis in a mouse model of haemophilia. *Nature* **475**, 217-21 (2011).
309. McDonald, J.I. *et al.* Reprogrammable CRISPR/Cas9-based system for inducing site-specific DNA methylation. *Biol Open* **5**, 866-74 (2016).
310. Vojta, A. *et al.* Repurposing the CRISPR-Cas9 system for targeted DNA methylation. *Nucleic Acids Res* **44**, 5615-28 (2016).
311. Liu, X.S. *et al.* Editing DNA Methylation in the Mammalian Genome. *Cell* **167**, 233-247 e17 (2016).
312. Mkannez, G. *et al.* DNA methylation of a PLPP3 MIR transposon-based enhancer promotes an osteogenic programme in calcific aortic valve disease. *Cardiovasc Res* **114**, 1525-1535 (2018).

313. Stepper, P. *et al.* Efficient targeted DNA methylation with chimeric dCas9-Dnmt3a-Dnmt3L methyltransferase. *Nucleic Acids Res* **45**, 1703-1713 (2017).
314. Lei, Y. *et al.* Targeted DNA methylation in vivo using an engineered dCas9-MQ1 fusion protein. *Nat Commun* **8**, 16026 (2017).
315. Meister, G.E., Chandrasegaran, S. & Ostermeier, M. An engineered split M.HhaI-zinc finger fusion lacks the intended methyltransferase specificity. *Biochem Biophys Res Commun* **377**, 226-30 (2008).
316. Galonska, C. *et al.* Genome-wide tracking of dCas9-methyltransferase footprints. *Nat Commun* **9**, 597 (2018).
317. Hofacker, D. *et al.* Engineering of Effector Domains for Targeted DNA Methylation with Reduced Off-Target Effects. *Int J Mol Sci* **21**(2020).
318. Slaska-Kiss, K. *et al.* Lowering DNA binding affinity of SssI DNA methyltransferase does not enhance the specificity of targeted DNA methylation in E. coli. *Sci Rep* **11**, 15226 (2021).
319. Huang, Y.H. *et al.* DNA epigenome editing using CRISPR-Cas SunTag-directed DNMT3A. *Genome Biol* **18**, 176 (2017).
320. Pflueger, C. *et al.* A modular dCas9-SunTag DNMT3A epigenome editing system overcomes pervasive off-target activity of direct fusion dCas9-DNMT3A constructs. *Genome Res* **28**, 1193-1206 (2018).
321. Xiong, T. *et al.* Targeted DNA methylation in human cells using engineered dCas9-methyltransferases. *Sci Rep* **7**, 6732 (2017).
322. Wang, Y.A. *et al.* DNA methyltransferase-3a interacts with p53 and represses p53-mediated gene expression. *Cancer Biol Ther* **4**, 1138-43 (2005).
323. Takahashi, Y. *et al.* Integration of CpG-free DNA induces de novo methylation of CpG islands in pluripotent stem cells. *Science* **356**, 503-508 (2017).
324. Cheng, S., Mayshar, Y. & Stelzer, Y. Induced epigenetic changes memorized across generations in mice. *Cell* **186**, 683-685 (2023).
325. Lei, H. *et al.* De novo DNA cytosine methyltransferase activities in mouse embryonic stem cells. *Development* **122**, 3195-205 (1996).
326. Fu, Y. *et al.* High-frequency off-target mutagenesis induced by CRISPR-Cas nucleases in human cells. *Nat Biotechnol* **31**, 822-6 (2013).
327. Papathanasiou, S. *et al.* Whole chromosome loss and genomic instability in mouse embryos after CRISPR-Cas9 genome editing. *Nat Commun* **12**, 5855 (2021).
328. Choudhury, S.R., Cui, Y., Lubecka, K., Stefanska, B. & Irudayaraj, J. CRISPR-dCas9 mediated TET1 targeting for selective DNA demethylation at BRCA1 promoter. *Oncotarget* **7**, 46545-46556 (2016).
329. Xu, X. *et al.* A CRISPR-based approach for targeted DNA demethylation. *Cell Discov* **2**, 16009 (2016).
330. Morita, S. *et al.* Targeted DNA demethylation in vivo using dCas9-peptide repeat and scFv-TET1 catalytic domain fusions. *Nat Biotechnol* **34**, 1060-1065 (2016).
331. Rienecker, K. PhD Thesis, Cardiff University (2018).
332. Xu, X. *et al.* High-fidelity CRISPR/Cas9- based gene-specific hydroxymethylation rescues gene expression and attenuates renal fibrosis. *Nat Commun* **9**, 3509 (2018).
333. Devesa-Guerra, I. *et al.* DNA Methylation Editing by CRISPR-guided Excision of 5-Methylcytosine. *J Mol Biol* **432**, 2204-2216 (2020).
334. Williams, K. *et al.* TET1 and hydroxymethylcytosine in transcription and DNA methylation fidelity. *Nature* **473**, 343-8 (2011).

335. Liu, X.S. *et al.* Rescue of Fragile X Syndrome Neurons by DNA Methylation Editing of the FMR1 Gene. *Cell* (2018).
336. Pfeifer, G.P., Szabo, P.E. & Song, J. Protein Interactions at Oxidized 5-Methylcytosine Bases. *J Mol Biol* (2019).
337. Baylin, S.B., Herman, J.G., Graff, J.R., Vertino, P.M. & Issa, J.P. Alterations in DNA methylation: a fundamental aspect of neoplasia. *Adv Cancer Res* **72**, 141-96 (1998).
338. Graff, J.R., Herman, J.G., Myohanen, S., Baylin, S.B. & Vertino, P.M. Mapping patterns of CpG island methylation in normal and neoplastic cells implicates both upstream and downstream regions in de novo methylation. *J Biol Chem* **272**, 22322-9 (1997).
339. Graff, J.R. *et al.* E-cadherin expression is silenced by DNA hypermethylation in human breast and prostate carcinomas. *Cancer Res* **55**, 5195-9 (1995).
340. Baylin, S.B. & Herman, J.G. DNA hypermethylation in tumorigenesis: epigenetics joins genetics. *Trends Genet* **16**, 168-74 (2000).
341. Richardson, B. DNA methylation and autoimmune disease. *Clin Immunol* **109**, 72-9 (2003).
342. Richardson, B. The interaction between environmental triggers and epigenetics in autoimmunity. *Clin Immunol* **192**, 1-5 (2018).
343. Heijmans, B.T. *et al.* Persistent epigenetic differences associated with prenatal exposure to famine in humans. *Proc Natl Acad Sci U S A* **105**, 17046-9 (2008).
344. Kim, J.Y. *et al.* Hypomethylation in MTNR1B: a novel epigenetic marker for atherosclerosis profiling using stenosis radiophenotype and blood inflammatory cells. *Clin Epigenetics* **15**, 11 (2023).
345. Ramos, R.B., Fabris, V., Lecke, S.B., Maturana, M.A. & Spritzer, P.M. Association between global leukocyte DNA methylation and cardiovascular risk in postmenopausal women. *BMC Med Genet* **17**, 71 (2016).
346. Har-Zahav, A. *et al.* The role of DNA demethylation in liver to pancreas transdifferentiation. *Stem Cell Res Ther* **13**, 476 (2022).
347. Waterland, R.A. & Michels, K.B. Epigenetic epidemiology of the developmental origins hypothesis. *Annu Rev Nutr* **27**, 363-88 (2007).
348. Rover, L.K. *et al.* PD-1 (PDCD1) Promoter Methylation Is a Prognostic Factor in Patients With Diffuse Lower-Grade Gliomas Harboring Isocitrate Dehydrogenase (IDH) Mutations. *EBioMedicine* **28**, 97-104 (2018).
349. Goltz, D. *et al.* Promoter methylation of the immune checkpoint receptor PD-1 (PDCD1) is an independent prognostic biomarker for biochemical recurrence-free survival in prostate cancer patients following radical prostatectomy. *Oncoimmunology* **5**, e1221555 (2016).
350. Robertson, K.D. DNA methylation and human disease. *Nat Rev Genet* **6**, 597-610 (2005).
351. Sutter, D. & Doerfler, W. Methylation of integrated adenovirus type 12 DNA sequences in transformed cells is inversely correlated with viral gene expression. *Proc Natl Acad Sci U S A* **77**, 253-6 (1980).
352. Laird, P.W. & Jaenisch, R. The role of DNA methylation in cancer genetic and epigenetics. *Annu Rev Genet* **30**, 441-64 (1996).
353. Labonte, B. *et al.* Genome-wide methylation changes in the brains of suicide completers. *Am J Psychiatry* **170**, 511-20 (2013).
354. Widschwendter, M. *et al.* Epigenetic stem cell signature in cancer. *Nat Genet* **39**, 157-8 (2007).

355. Ohm, J.E. *et al.* A stem cell-like chromatin pattern may predispose tumor suppressor genes to DNA hypermethylation and heritable silencing. *Nat Genet* **39**, 237-42 (2007).
356. Gal-Yam, E.N. *et al.* Frequent switching of Polycomb repressive marks and DNA hypermethylation in the PC3 prostate cancer cell line. *Proc Natl Acad Sci U S A* **105**, 12979-84 (2008).
357. Schlesinger, Y. *et al.* Polycomb-mediated methylation on Lys27 of histone H3 pre-marks genes for de novo methylation in cancer. *Nat Genet* **39**, 232-6 (2007).
358. Szyf, M., Rouleau, J., Theberge, J. & Bozovic, V. Induction of myogenic differentiation by an expression vector encoding the DNA methyltransferase cDNA sequence in the antisense orientation. *J Biol Chem* **267**, 12831-6 (1992).
359. Schmidt, T., Leha, A. & Salinas-Riester, G. Treatment of prostate cancer cells with S-adenosylmethionine leads to genome-wide alterations in transcription profiles. *Gene* **595**, 161-167 (2016).
360. Frauer, C. *et al.* Recognition of 5-hydroxymethylcytosine by the Uhrf1 SRA domain. *PLoS One* **6**, e21306 (2011).
361. Hashimoto, H. *et al.* Wilms tumor protein recognizes 5-carboxylcytosine within a specific DNA sequence. *Genes Dev* **28**, 2304-13 (2014).
362. Jin, S.G. *et al.* Tet3 Reads 5-Carboxylcytosine through Its CXXC Domain and Is a Potential Guardian against Neurodegeneration. *Cell Rep* **14**, 493-505 (2016).
363. Schneider, M. *et al.* Systematic analysis of the binding behaviour of UHRF1 towards different methyl- and carboxylcytosine modification patterns at CpG dyads. *PLoS One* **15**, e0229144 (2020).
364. Spruijt, C.G. *et al.* Dynamic readers for 5-(hydroxy)methylcytosine and its oxidized derivatives. *Cell* **152**, 1146-59 (2013).
365. Wang, L. *et al.* Molecular basis for 5-carboxycytosine recognition by RNA polymerase II elongation complex. *Nature* **523**, 621-5 (2015).
366. Raiber, E.A. *et al.* 5-Formylcytosine alters the structure of the DNA double helix. *Nat Struct Mol Biol* **22**, 44-49 (2015).
367. Deplus, R. *et al.* TET2 and TET3 regulate GlcNAcylation and H3K4 methylation through OGT and SET1/COMPASS. *EMBO J* **32**, 645-55 (2013).
368. Hrit, J., Li, C., Martin, E.A., Goll, M. & Panning, B. OGT binds a conserved C-terminal domain of TET1 to regulate TET1 activity and function in development. *bioRxiv* (2017).
369. Polumuri, S.K. *et al.* Transcriptional regulation of murine IL-33 by TLR and non-TLR agonists. *J Immunol* **189**, 50-60 (2012).
370. Talabot-Ayer, D. *et al.* The mouse interleukin (Il)33 gene is expressed in a cell type- and stimulus-dependent manner from two alternative promoters. *J Leukoc Biol* **91**, 119-25 (2012).
371. Ikeda, K., Stuehler, T. & Meisterernst, M. The H1 and H2 regions of the activation domain of herpes simplex virion protein 16 stimulate transcription through distinct molecular mechanisms. *Genes Cells* **7**, 49-58 (2002).
372. Adli, M. The CRISPR tool kit for genome editing and beyond. *Nature communications* **9**, 1911 (2018).
373. Herman, J.G. *et al.* Inactivation of the CDKN2/p16/MTS1 gene is frequently associated with aberrant DNA methylation in all common human cancers. *Cancer Res* **55**, 4525-30 (1995).
374. Vilkaitis, G., Suetake, I., Klimasauskas, S. & Tajima, S. Processive methylation of hemimethylated CpG sites by mouse Dnmt1 DNA methyltransferase. *J Biol Chem* **280**, 64-72 (2005).

375. Nishimasu, H. *et al.* Crystal structure of Cas9 in complex with guide RNA and target DNA. *Cell* **156**, 935-49 (2014).
376. Zhu, X. *et al.* Cryo-EM structures reveal coordinated domain motions that govern DNA cleavage by Cas9. *Nature structural & molecular biology* **26**, 679-685 (2019).
377. Leonhardt, H., Page, A.W., Weier, H.U. & Bestor, T.H. A targeting sequence directs DNA methyltransferase to sites of DNA replication in mammalian nuclei. *Cell* **71**, 865-73 (1992).
378. Ran, F.A. *et al.* Genome engineering using the CRISPR-Cas9 system. *Nature protocols* **8**, 2281-308 (2013).
379. Jinek, M. *et al.* RNA-programmed genome editing in human cells. *Elife* **2**, e00471 (2013).
380. Yoshioka, S., Fujii, W., Ogawa, T., Sugiura, K. & Naito, K. Development of a mono-promoter-driven CRISPR/Cas9 system in mammalian cells. *Scientific reports* **5**, 18341 (2015).
381. Fujii, W., Kawasaki, K., Sugiura, K. & Naito, K. Efficient generation of large-scale genome-modified mice using gRNA and CAS9 endonuclease. *Nucleic acids research* **41**, e187 (2013).
382. Ren, X. *et al.* Enhanced specificity and efficiency of the CRISPR/Cas9 system with optimized sgRNA parameters in Drosophila. *Cell reports* **9**, 1151-62 (2014).
383. Yang, X. *et al.* Synergistic activation of functional estrogen receptor (ER)-alpha by DNA methyltransferase and histone deacetylase inhibition in human ER-alpha-negative breast cancer cells. *Cancer Res* **61**, 7025-9 (2001).
384. Sponheim, J. *et al.* Inflammatory bowel disease-associated interleukin-33 is preferentially expressed in ulceration-associated myofibroblasts. *Am J Pathol* **177**, 2804-15 (2010).
385. Aranyi, T. *et al.* Systemic epigenetic response to recombinant lentiviral vectors independent of proviral integration. *Epigenetics Chromatin* **9**, 29 (2016).
386. Yamagata, Y. *et al.* Lentiviral transduction of CD34(+) cells induces genome-wide epigenetic modifications. *PLoS One* **7**, e48943 (2012).
387. Stemmer, M., Thumberger, T., Del Sol Keyer, M., Wittbrodt, J. & Mateo, J.L. CCTop: An Intuitive, Flexible and Reliable CRISPR/Cas9 Target Prediction Tool. *PloS one* **10**, e0124633 (2015).
388. Pliatsika, V. & Rigoutsos, I. "Off-Spotter": very fast and exhaustive enumeration of genomic lookalikes for designing CRISPR/Cas guide RNAs. *Biology direct* **10**, 4 (2015).
389. Bae, S., Park, J. & Kim, J.S. Cas-OFFinder: a fast and versatile algorithm that searches for potential off-target sites of Cas9 RNA-guided endonucleases. *Bioinformatics* **30**, 1473-5 (2014).
390. Wu, X. *et al.* Genome-wide binding of the CRISPR endonuclease Cas9 in mammalian cells. *Nature biotechnology* **32**, 670-6 (2014).
391. Zhang, X.-H., Tee, L.Y., Wang, X.-G., Huang, Q.-S. & Yang, S.-H. Off-target Effects in CRISPR/Cas9-mediated Genome Engineering. *Molecular therapy Nucleic acids* **4**, e264 (2015).
392. Tanasijevic, B. *et al.* Progressive accumulation of epigenetic heterogeneity during human ES cell culture. *Epigenetics* **4**, 330-8 (2009).
393. Bushman, F. *et al.* Genome-wide analysis of retroviral DNA integration. *Nature reviews Microbiology* **3**, 848-58 (2005).
394. Lu, F., Liu, Y., Jiang, L., Yamaguchi, S. & Zhang, Y. Role of Tet proteins in enhancer activity and telomere elongation. *Genes & development* **28**, 2103-19 (2014).

395. Wiehle, L. *et al.* Tet1 and Tet2 Protect DNA Methylation Canyons against Hypermethylation. *Molecular and cellular biology* **36**, 452-61 (2016).
396. Serandour, A.A. *et al.* Dynamic hydroxymethylation of deoxyribonucleic acid marks differentiation-associated enhancers. *Nucleic acids research* **40**, 8255-65 (2012).
397. Charlton, J. *et al.* TETs compete with DNMT3 activity in pluripotent cells at thousands of methylated somatic enhancers. *Nature genetics* **52**, 819-827 (2020).
398. Hon, G.C. *et al.* 5mC oxidation by Tet2 modulates enhancer activity and timing of transcriptome reprogramming during differentiation. *Molecular cell* **56**, 286-297 (2014).
399. Wang, L. *et al.* TET2 coactivates gene expression through demethylation of enhancers. *Science advances* **4**, eaau6986 (2018).
400. Lizio, M. *et al.* Gateways to the FANTOM5 promoter level mammalian expression atlas. *Genome biology* **16**, 22 (2015).
401. Zou, Z. *et al.* Maspin, a serpin with tumor-suppressing activity in human mammary epithelial cells. *Science (New York, N Y)* **263**, 526-9 (1994).
402. Seftor, R.E. *et al.* maspin suppresses the invasive phenotype of human breast carcinoma. *Cancer research* **58**, 5681-5 (1998).
403. Zhang, M., Volpert, O., Shi, Y.H. & Bouck, N. Maspin is an angiogenesis inhibitor. *Nature medicine* **6**, 196-9 (2000).
404. Sheng, S. *et al.* Maspin acts at the cell membrane to inhibit invasion and motility of mammary and prostatic cancer cells. *Proceedings of the National Academy of Sciences of the United States of America* **93**, 11669-74 (1996).
405. Beltran, A. *et al.* Re-activation of a dormant tumor suppressor gene maspin by designed transcription factors. *Oncogene* **26**, 2791-8 (2007).
406. Futscher, B.W. *et al.* Role for DNA methylation in the control of cell type specific maspin expression. *Nature genetics* **31**, 175-9 (2002).
407. Sato, N., Fukushima, N., Matsubayashi, H. & Goggins, M. Identification of maspin and S100P as novel hypomethylation targets in pancreatic cancer using global gene expression profiling. *Oncogene* **23**, 1531-8 (2004).
408. Domann, F.E., Rice, J.C., Hendrix, M.J. & Futscher, B.W. Epigenetic silencing of maspin gene expression in human breast cancers. *International journal of cancer* **85**, 805-10 (2000).
409. Wu, Y., Alvarez, M., Slamon, D.J., Koeffler, P. & Vadgama, J.V. Caspase 8 and maspin are downregulated in breast cancer cells due to CpG site promoter methylation. *BMC cancer* **10**, 32 (2010).
410. Wada, K., Maesawa, C., Akasaka, T. & Masuda, T. Aberrant expression of the maspin gene associated with epigenetic modification in melanoma cells. *The Journal of investigative dermatology* **122**, 805-11 (2004).
411. Tsytsykova, A.V. *et al.* Activation-dependent intrachromosomal interactions formed by the TNF gene promoter and two distal enhancers. *Proc Natl Acad Sci U S A* **104**, 16850-5 (2007).
412. O'Donnell, W.T. & Warren, S.T. A decade of molecular studies of fragile X syndrome. *Annual Review of Neuroscience* **25**, 315-338 (2002).
413. Coffee, B., Zhang, F., Warren, S.T. & Reines, D. Acetylated histones are associated with FMR1 in normal but not fragile X-syndrome cells. *Nature genetics* **22**, 98-101 (1999).
414. Chiurazzi, P., Pomponi, M.G., Willemsen, R., Oostra, B.A. & Neri, G. In vitro reactivation of the FMR1 gene involved in fragile X syndrome. *Human molecular genetics* **7**, 109-13 (1998).

415. Vershkov, D. *et al.* FMR1 Reactivating Treatments in Fragile X iPSC-Derived Neural Progenitors InVitro and InVivo. *Cell reports* **26**, 2531-2539.e4 (2019).
416. Biacsi, R., Kumari, D. & Usdin, K. SIRT1 inhibition alleviates gene silencing in Fragile X mental retardation syndrome. *PLoS genetics* **4**, e1000017 (2008).
417. Sheridan, S.D. *et al.* Epigenetic Characterization of the FMR1 Gene and Aberrant Neurodevelopment in Human Induced Pluripotent Stem Cell Models of Fragile X Syndrome. *Plos One* **6**(2011).
418. Liu, X.S. *et al.* Rescue of Fragile X Syndrome Neurons by DNA Methylation Editing of the FMR1 Gene. *Cell* **172**, 979-+ (2018).
419. Allen, B., Pezone, A., Porcellini, A., Muller, M.T. & Masternak, M.M. Non-homologous end joining induced alterations in DNA methylation: A source of permanent epigenetic change. *Oncotarget* **8**, 40359-40372 (2017).
420. Roadmap Epigenomics, C. *et al.* Integrative analysis of 111 reference human epigenomes. *Nature* **518**, 317-30 (2015).
421. Robertson, K.D. & Wolffe, A.P. DNA methylation in health and disease. *Nat Rev Genet* **1**, 11-9 (2000).
422. Bergman, Y. & Cedar, H. DNA methylation dynamics in health and disease. *Nat Struct Mol Biol* **20**, 274-81 (2013).
423. Rountree, M.R., Bachman, K.E., Herman, J.G. & Baylin, S.B. DNA methylation, chromatin inheritance, and cancer. *Oncogene* **20**, 3156-65. (2001).
424. Shames, D.S., Minna, J.D. & Gazdar, A.F. DNA methylation in health, disease, and cancer. *Curr Mol Med* **7**, 85-102 (2007).
425. Fuchikami, M. *et al.* DNA methylation profiles of the brain-derived neurotrophic factor (BDNF) gene as a potent diagnostic biomarker in major depression. *PLoS One* **6**, e23881 (2011).
426. Kordi-Tamandani, D.M., Sahranavard, R. & Torkamanzehi, A. DNA methylation and expression profiles of the brain-derived neurotrophic factor (BDNF) and dopamine transporter (DAT1) genes in patients with schizophrenia. *Mol Biol Rep* **39**, 10889-93 (2012).
427. Kuratomi, G. *et al.* Aberrant DNA methylation associated with bipolar disorder identified from discordant monozygotic twins. *Mol Psychiatry* **13**, 429-41 (2008).
428. van der Ploeg, L.H. & Flavell, R.A. DNA methylation in the human gamma delta beta-globin locus in erythroid and nonerythroid tissues. *Cell* **19**, 947-58 (1980).
429. Vardimon, L., Kuhlmann, I., Doerfler, W. & Cedar, H. Methylation of adenovirus genes in transformed cells and in vitro: influence on the regulation of gene expression? *Eur J Cell Biol* **25**, 13-5 (1981).
430. Stein, R., Razin, A. & Cedar, H. In vitro methylation of the hamster adenine phosphoribosyltransferase gene inhibits its expression in mouse L cells. *Proc Natl Acad Sci U S A* **79**, 3418-22 (1982).
431. Razin, A. & Szyf, M. DNA methylation patterns. Formation and function. *Biochim Biophys Acta* **782**, 331-42 (1984).
432. Walsh, C.P., Chaillet, J.R. & Bestor, T.H. Transcription of IAP endogenous retroviruses is constrained by cytosine methylation [letter]. *Nat Genet* **20**, 116-7 (1998).
433. Jones, P.A. Effects of 5-azacytidine and its 2'-deoxyderivative on cell differentiation and DNA methylation. *Pharmacol Ther* **28**, 17-27 (1985).
434. Rountree, M.R., Bachman, K.E. & Baylin, S.B. DNMT1 binds HDAC2 and a new co-repressor, DMAP1, to form a complex at replication foci. *Nat Genet* **25**, 269-77 (2000).

435. Fuks, F., Burgers, W.A., Brehm, A., Hughes-Davies, L. & Kouzarides, T. DNA methyltransferase Dnmt1 associates with histone deacetylase activity. *Nat Genet* **24**, 88-91 (2000).
436. Bai, S. *et al.* DNA methyltransferase 3b regulates nerve growth factor-induced differentiation of PC12 cells by recruiting histone deacetylase 2. *Mol Cell Biol* **25**, 751-66 (2005).
437. Massart, R., Suderman, M., Mongrain, V. & Szyf, M. DNA methylation and transcription onset in the brain. *Epigenomics* (2017).
438. Suzuki, T. *et al.* A screening system to identify transcription factors that induce binding site-directed DNA demethylation. *Epigenetics Chromatin* **10**, 60 (2017).
439. Kirillov, A. *et al.* A role for nuclear NF-kappaB in B-cell-specific demethylation of the Igkappa locus. *Nat Genet* **13**, 435-41 (1996).
440. Mayran, A. & Drouin, J. Pioneer transcription factors shape the epigenetic landscape. *J Biol Chem* **293**, 13795-13804 (2018).
441. Pfaffeneder, T. *et al.* Tet oxidizes thymine to 5-hydroxymethyluracil in mouse embryonic stem cell DNA. *Nat Chem Biol* **10**, 574-81 (2014).
442. Devesa-Guerra, I. *et al.* DNA methylation editing by CRISPR-guided excision of 5-methylcytosine. *J Mol Biol* (2020).
443. Vella, P. *et al.* Tet proteins connect the O-linked N-acetylglucosamine transferase Ogt to chromatin in embryonic stem cells. *Mol Cell* **49**, 645-56 (2013).
444. Dixon, J.R. *et al.* Topological domains in mammalian genomes identified by analysis of chromatin interactions. *Nature* **485**, 376-80 (2012).
445. Fang, S. *et al.* Tet inactivation disrupts YY1 binding and long-range chromatin interactions during embryonic heart development. *Nature communications* **10**, 4297 (2019).
446. Senner, C.E. *et al.* TET1 and 5-Hydroxymethylation Preserve the Stem Cell State of Mouse Trophoblast. *Stem cell reports* (2020).
447. Maeder, M.L. *et al.* Targeted DNA demethylation and activation of endogenous genes using programmable TALE-TET1 fusion proteins. *Nat Biotechnol* **31**, 1137-42 (2013).
448. Hata, K. & Sakaki, Y. Identification of critical CpG sites for repression of L1 transcription by DNA methylation. *Gene* **189**, 227-34 (1997).
449. Cheishvili, D. *et al.* DNA methylation controls unmethylated transcription start sites in the genome in trans. *Epigenomics* (2017).
450. Ficz, G. *et al.* FGF signaling inhibition in ESCs drives rapid genome-wide demethylation to the epigenetic ground state of pluripotency. *Cell Stem Cell* **13**, 351-9 (2013).
451. Mali, P. *et al.* RNA-guided human genome engineering via Cas9. *Science (New York, N Y)* **339**, 823-6 (2013).
452. Sanjana, N.E., Shalem, O. & Zhang, F. Improved vectors and genome-wide libraries for CRISPR screening. *Nat Methods* **11**, 783-784 (2014).
453. Li, L.C. & Dahiya, R. MethPrimer: designing primers for methylation PCRs. *Bioinformatics* **18**, 1427-31 (2002).
454. Untergasser, A. *et al.* Primer3--new capabilities and interfaces. *Nucleic Acids Res* **40**, e115 (2012).
455. Krueger, F. Trim galore: A wrapper tool around Cutadapt and FastQC to consistently apply quality and adapter trimming to FastQ files. (2015).
456. Langmead, B. & Salzberg, S.L. Fast gapped-read alignment with Bowtie 2. *Nat Methods* **9**, 357-9 (2012).

457. Zhang, Y. *et al.* Model-based analysis of ChIP-Seq (MACS). *Genome Biol* **9**, R137 (2008).
458. Krueger, F. & Andrews, S.R. Bismark: a flexible aligner and methylation caller for Bisulfite-Seq applications. *Bioinformatics (Oxford, England)* **27**, 1571-2 (2011).
459. Akalin, A. *et al.* methylKit: a comprehensive R package for the analysis of genome-wide DNA methylation profiles. *Genome biology* **13**, R87 (2012).
460. Turner, S.D. qqman: an R package for visualizing GWAS results using Q-Q and manhattan plots. *BioRxiv* (2014).
461. Ho, D.W.H., Sze, K.M.F. & Ng, I.O.L. Virus-Clip: a fast and memory-efficient viral integration site detection tool at single-base resolution with annotation capability. *Oncotarget* **6**, 20959-63 (2015).
462. Li, H. *et al.* The Sequence Alignment/Map format and SAMtools. *Bioinformatics (Oxford, England)* **25**, 2078-9 (2009).
463. Altschul, S.F., Gish, W., Miller, W., Myers, E.W. & Lipman, D.J. Basic local alignment search tool. *Journal of molecular biology* **215**, 403-10 (1990).
464. Quinlan, A.R. & Hall, I.M. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics (Oxford, England)* **26**, 841-2 (2010).
465. Moore, L.D., Le, T. & Fan, G. DNA methylation and its basic function. *Neuropsychopharmacology* **38**, 23-38 (2013).
466. Day, J.J. *et al.* DNA methylation regulates associative reward learning. *Nat Neurosci* **16**, 1445-52 (2013).
467. Mellen, M., Ayata, P., Dewell, S., Kriaucionis, S. & Heintz, N. MeCP2 binds to 5hmC enriched within active genes and accessible chromatin in the nervous system. *Cell* **151**, 1417-30 (2012).
468. Cramer, J.M. *et al.* Probing the dynamic distribution of bound states for methylcytosine-binding domains on DNA. *J Biol Chem* **289**, 1294-302 (2014).
469. Mellen, M., Ayata, P. & Heintz, N. 5-hydroxymethylcytosine accumulation in postmitotic neurons results in functional demethylation of expressed genes. *Proc Natl Acad Sci U S A* **114**, E7812-E7821 (2017).
470. Hashimoto, H. *et al.* Recognition and potential mechanisms for replication and erasure of cytosine hydroxymethylation. *Nucleic Acids Res* **40**, 4841-9 (2012).
471. James, S.J., Shpyleva, S., Melnyk, S., Pavliv, O. & Pogribny, I.P. Elevated 5-hydroxymethylcytosine in the Engrailed-2 (EN-2) promoter is associated with increased gene expression and decreased MeCP2 binding in autism cerebellum. *Transl Psychiatry* **4**, e460 (2014).
472. Cortellino, S. *et al.* Thymine DNA glycosylase is essential for active DNA demethylation by linked deamination-base excision repair. *Cell* **146**, 67-79 (2011).
473. Schutsky, E.K. *et al.* Nondestructive, base-resolution sequencing of 5-hydroxymethylcytosine using a DNA deaminase. *Nat Biotechnol* (2018).
474. Zhang, L. *et al.* Thymine DNA glycosylase specifically recognizes 5-carboxylcytosine-modified DNA. *Nat Chem Biol* **8**, 328-30 (2012).
475. Hardeland, U., Steinacher, R., Jiricny, J. & Schar, P. Modification of the human thymine-DNA glycosylase by ubiquitin-like proteins facilitates enzymatic turnover. *EMBO J* **21**, 1456-64 (2002).
476. Steinacher, R. & Schar, P. Functionality of human thymine DNA glycosylase requires SUMO-regulated changes in protein conformation. *Curr Biol* **15**, 616-23 (2005).

477. D'Alessio, A.C., Weaver, I.C. & Szyf, M. Acetylation-induced transcription is required for active DNA demethylation in methylation-silenced genes. *Mol Cell Biol* **27**, 7462-74 (2007).
478. Kim, D.V. *et al.* Mild phenotype of knockouts of the major apurinic/apyrimidinic endonuclease APEX1 in a non-cancer human cell line. *PLoS One* **16**, e0257473 (2021).
479. Paul, S. *et al.* Co-expression networks in generation of induced pluripotent stem cells. *Biol Open* **5**, 300-10 (2016).
480. Muhl, L. *et al.* Single-cell analysis uncovers fibroblast heterogeneity and criteria for fibroblast and mural cell identification and discrimination. *Nat Commun* **11**, 3953 (2020).
481. Torres, L. *et al.* Liver-specific methionine adenosyltransferase MAT1A gene expression is associated with a specific pattern of promoter methylation and histone acetylation: implications for MAT1A silencing during transformation. *FASEB J* **14**, 95-102 (2000).
482. Razonable, R.R. Innate immune genetic profile to predict infection risk and outcome after liver transplant. *Hepatology* **52**, 814-17 (2010).
483. Cardot, P., Chambaz, J., Cladaras, C. & Zannis, V.I. Regulation of the human ApoA-II gene by the synergistic action of factors binding to the proximal and distal regulatory elements. *J Biol Chem* **266**, 24460-70 (1991).
484. Scheurer, B., Rittner, C. & Schneider, P.M. Expression of the human complement C8 subunits is independently regulated by interleukin 1 beta, interleukin 6, and interferon gamma. *Immunopharmacology* **38**, 167-75 (1997).
485. Mandal, S. & Davie, J.R. An integrated analysis of genes and pathways exhibiting metabolic differences between estrogen receptor positive breast cancer cells. *BMC Cancer* **7**, 181 (2007).
486. Meitinger, F. *et al.* TRIM37 controls cancer-specific vulnerability to PLK4 inhibition. *Nature* **585**, 440-446 (2020).
487. Basta, J. & Rauchman, M. The nucleosome remodeling and deacetylase complex in development and disease. *Transl Res* **165**, 36-47 (2015).
488. Sarshad, A. *et al.* Nuclear myosin 1c facilitates the chromatin modifications required to activate rRNA gene transcription and cell cycle progression. *PLoS Genet* **9**, e1003397 (2013).
489. Cavellan, E., Asp, P., Percipalle, P. & Farrants, A.K. The WSTF-SNF2h chromatin remodeling complex interacts with several nuclear proteins in transcription. *J Biol Chem* **281**, 16264-71 (2006).
490. Brown, S.E. & Szyf, M. Epigenetic programming of the rRNA promoter by MBD3. *Mol Cell Biol* **27**, 4938-52 (2007).
491. Xie, W. *et al.* The chromatin remodeling complex NuRD establishes the poised state of rRNA genes characterized by bivalent histone modifications and altered nucleosome positions. *Proc Natl Acad Sci U S A* **109**, 8161-6 (2012).
492. Gunther, K. *et al.* Differential roles for MBD2 and MBD3 at methylated CpG islands, active promoters and binding to exon sequences. *Nucleic Acids Res* **41**, 3010-21 (2013).
493. Le Guezennec, X. *et al.* MBD2/NuRD and MBD3/NuRD, two distinct complexes with different biochemical and functional properties. *Mol Cell Biol* **26**, 843-51 (2006).
494. Kaji, K. *et al.* The NuRD component Mbd3 is required for pluripotency of embryonic stem cells. *Nat Cell Biol* **8**, 285-92 (2006).
495. Liu, H.W., Banerjee, T., Guan, X., Freitas, M.A. & Parvin, J.D. The chromatin scaffold protein SAFB1 localizes SUMO-1 to the promoters of ribosomal protein genes to facilitate transcription initiation and splicing. *Nucleic Acids Res* **43**, 3605-13 (2015).

496. Nayler, O. *et al.* SAF-B protein couples transcription and pre-mRNA splicing to SAR/MAR elements. *Nucleic Acids Res* **26**, 3542-9 (1998).
497. Liu, J. *et al.* Chromatin modifier MTA1 regulates mitotic transition and tumorigenesis by orchestrating mitotic mRNA processing. *Nat Commun* **11**, 4455 (2020).
498. Oughtred, R. *et al.* The BioGRID database: A comprehensive biomedical resource of curated protein, genetic, and chemical interactions. *Protein Sci* **30**, 187-200 (2021).
499. Leontiou, C.A. *et al.* Bisulfite Conversion of DNA: Performance Comparison of Different Kits and Methylation Quantitation of Epigenetic Biomarkers that Have the Potential to Be Used in Non-Invasive Prenatal Testing. *PLoS One* **10**, e0135058 (2015).
500. Consortium, E.P. An integrated encyclopedia of DNA elements in the human genome. *Nature* **489**, 57-74 (2012).
501. Ninova, M., Fejes Toth, K. & Aravin, A.A. The control of gene expression and cell identity by H3K9 trimethylation. *Development* **146**(2019).
502. Cui, X.L. *et al.* A human tissue map of 5-hydroxymethylcytosines exhibits tissue specificity through gene and enhancer modulation. *Nat Commun* **11**, 6161 (2020).
503. Gibas, P. *et al.* Precise genomic mapping of 5-hydroxymethylcytosine via covalent tether-directed sequencing. *PLoS Biol* **18**, e3000684 (2020).
504. Sethi, S. *et al.* A holistic view of mouse enhancer architectures reveals analogous pleiotropic effects and correlation with human disease. *BMC Genomics* **21**, 754 (2020).
505. Li, Y.E. *et al.* An atlas of gene regulatory elements in adult mouse cerebrum. *Nature* **598**, 129-136 (2021).
506. Muller, U., Bauer, C., Siegl, M., Rottach, A. & Leonhardt, H. TET-mediated oxidation of methylcytosine causes TDG or NEIL glycosylase dependent gene reactivation. *Nucleic Acids Res* **42**, 8592-604 (2014).
507. Storebjerg, T.M. *et al.* Dysregulation and prognostic potential of 5-methylcytosine (5mC), 5-hydroxymethylcytosine (5hmC), 5-formylcytosine (5fC), and 5-carboxylcytosine (5caC) levels in prostate cancer. *Clin Epigenetics* **10**, 105 (2018).
508. Neary, J.L., Perez, S.M., Peterson, K., Lodge, D.J. & Carless, M.A. Comparative analysis of MBD-seq and MeDIP-seq and estimation of gene expression changes in a rodent model of schizophrenia. *Genomics* **109**, 204-213 (2017).
509. Buchmuller, B.C., Kosel, B. & Summerer, D. Complete Profiling of Methyl-CpG-Binding Domains for Combinations of Cytosine Modifications at CpG Dinucleotides Reveals Differential Read-out in Normal and Rett-Associated States. *Sci Rep* **10**, 4053 (2020).
510. Kungulovski, G. *et al.* Targeted epigenome editing of an endogenous locus with chromatin modifiers is not stably maintained. *Epigenetics Chromatin* **8**, 12 (2015).
511. O'Geen, H., Tomkova, M., Combs, J.A., Tilley, E.K. & Segal, D.J. Determinants of heritable gene silencing for KRAB-dCas9 + DNMT3 and Ezh2-dCas9 + DNMT3 hit-and-run epigenome editing. *Nucleic Acids Res* **50**, 3239-3253 (2022).
512. Grosser, C., Wagner, N., Grothaus, K. & Horsthemke, B. Altering TET dioxygenase levels within physiological range affects DNA methylation dynamics of HEK293 cells. *Epigenetics* **10**, 819-33 (2015).
513. Onodera, A. *et al.* Roles of TET and TDG in DNA demethylation in proliferating and non-proliferating immune cells. *Genome Biol* **22**, 186 (2021).
514. Shimbo, T. *et al.* MBD3 localizes at promoters, gene bodies and enhancers of active genes. *PLoS Genet* **9**, e1004028 (2013).
515. Wang, Z. *et al.* Genome-wide mapping of HATs and HDACs reveals distinct functions in active and inactive genes. *Cell* **138**, 1019-31 (2009).

516. Gryder, B.E. *et al.* Chemical genomics reveals histone deacetylases are required for core regulatory transcription. *Nat Commun* **10**, 3004 (2019).
517. Brown, S.E., Suderman, M.J., Hallett, M. & Szyf, M. DNA demethylation induced by the methyl-CpG-binding domain protein MBD3. *Gene* **420**, 99-106 (2008).
518. Hendrich, B., Guy, J., Ramsahoye, B., Wilson, V.A. & Bird, A. Closely related proteins MBD2 and MBD3 play distinctive but interacting roles in mouse development. *Genes Dev* **15**, 710-23 (2001).
519. Kaji, K., Nichols, J. & Hendrich, B. Mbd3, a component of the NuRD co-repressor complex, is required for development of pluripotent cells. *Development* **134**, 1123-32 (2007).
520. Mikkelsen, T.S. *et al.* Dissecting direct reprogramming through integrative genomic analysis. *Nature* **454**, 49-55 (2008).
521. Bagci, H. & Fisher, A.G. DNA demethylation in pluripotency and reprogramming: the role of tet proteins and cell division. *Cell Stem Cell* **13**, 265-9 (2013).
522. Iurlaro, M. *et al.* In vivo genome-wide profiling reveals a tissue-specific role for 5-formylcytosine. *Genome Biol* **17**, 141 (2016).
523. Husain, A., Begum, N.A., Kobayashi, M. & Honjo, T. Native Co-immunoprecipitation Assay to Identify Interacting Partners of Chromatin-associated Proteins in Mammalian Cells. *Bio Protoc* **10**, e3837 (2020).
524. Li, H. *et al.* The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25**, 2078-9 (2009).
525. Ramirez, F. *et al.* deepTools2: a next generation web server for deep-sequencing data analysis. *Nucleic Acids Res* **44**, W160-5 (2016).
526. Quinlan, A.R. & Hall, I.M. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics* **26**, 841-2 (2010).
527. Kent, W.J., Zweig, A.S., Barber, G., Hinrichs, A.S. & Karolchik, D. BigWig and BigBed: enabling browsing of large distributed datasets. *Bioinformatics* **26**, 2204-7 (2010).
528. Gel, B. & Serra, E. karyoploteR: an R/Bioconductor package to plot customizable genomes displaying arbitrary data. *Bioinformatics* **33**, 3088-3090 (2017).
529. Heinz, S. *et al.* Simple combinations of lineage-determining transcription factors prime cis-regulatory elements required for macrophage and B cell identities. *Mol Cell* **38**, 576-89 (2010).
530. Ge, S.X., Jung, D. & Yao, R. ShinyGO: a graphical gene-set enrichment tool for animals and plants. *Bioinformatics* **36**, 2628-2629 (2020).
531. Akalin, A. *et al.* methylKit: a comprehensive R package for the analysis of genome-wide DNA methylation profiles. *Genome Biol* **13**, R87 (2012).
532. Pogribny, I.P., Pogribna, M., Christman, J.K. & James, S.J. Single-site methylation within the p53 promoter region reduces gene expression in a reporter gene construct: possible in vivo relevance during tumorigenesis. *Cancer Res* **60**, 588-94 (2000).
533. Yang, L. *et al.* Methylation of a CGATA element inhibits binding and regulation by GATA-1. *Nat Commun* **11**, 2560 (2020).
534. Furst, R.W., Kliem, H., Meyer, H.H. & Ulbrich, S.E. A differentially methylated single CpG-site is correlated with estrogen receptor alpha transcription. *J Steroid Biochem Mol Biol* **130**, 96-104 (2012).
535. Nile, C.J., Read, R.C., Akil, M., Duff, G.W. & Wilson, A.G. Methylation status of a single CpG site in the IL6 promoter is related to IL6 messenger RNA levels and rheumatoid arthritis. *Arthritis Rheum* **58**, 2686-93 (2008).

536. Bordagaray, M.J. *et al.* CpG Single-Site Methylation Regulates TLR2 Expression in Proinflammatory PBMCs From Apical Periodontitis Individuals. *Front Immunol* **13**, 861665 (2022).
537. Patil, V., Ward, R.L. & Hesson, L.B. The evidence for functional non-CpG methylation in mammalian cells. *Epigenetics* **9**, 823-8 (2014).
538. Jang, H.S., Shin, W.J., Lee, J.E. & Do, J.T. CpG and Non-CpG Methylation in Epigenetic Gene Regulation and Brain Function. *Genes (Basel)* **8**(2017).
539. de Mendoza, A. *et al.* The emergence of the brain non-CpG methylation system in vertebrates. *Nat Ecol Evol* **5**, 369-378 (2021).
540. Wang, D. *et al.* CRISPR Screening of CAR T Cells and Cancer Stem Cells Reveals Critical Dependencies for Cell-Based Therapies. *Cancer Discov* **11**, 1192-1211 (2021).
541. Ye, L. *et al.* A genome-scale gain-of-function CRISPR screen in CD8 T cells identifies proline metabolism as a means to enhance CAR-T therapy. *Cell Metab* **34**, 595-614 e14 (2022).
542. Meivar-Levy, I. & Ferber, S. Liver to Pancreas Transdifferentiation. *Curr Diab Rep* **19**, 76 (2019).
543. Kim, S., Kim, D., Cho, S.W., Kim, J. & Kim, J.S. Highly efficient RNA-guided genome editing in human cells via delivery of purified Cas9 ribonucleoproteins. *Genome Res* **24**, 1012-9 (2014).
544. Schumann, K. *et al.* Generation of knock-in primary human T cells using Cas9 ribonucleoproteins. *Proc Natl Acad Sci U S A* **112**, 10437-42 (2015).
545. Li, Y., Glass, Z., Huang, M., Chen, Z.Y. & Xu, Q. Ex vivo cell-based CRISPR/Cas9 genome editing for therapeutic applications. *Biomaterials* **234**, 119711 (2020).
546. Yu, W. & Wu, Z. Use of AAV Vectors for CRISPR-Mediated In Vivo Genome Editing in the Retina. *Methods Mol Biol* **1950**, 123-139 (2019).
547. Hana, S. *et al.* Highly efficient neuronal gene knockout in vivo by CRISPR-Cas9 via neonatal intracerebroventricular injection of AAV in mice. *Gene Ther* **28**, 646-658 (2021).
548. Kreitz, J. *et al.* Programmable protein delivery with a bacterial contractile injection system. *Nature* (2023).
549. Sago, C.D. *et al.* Augmented lipid-nanoparticle-mediated in vivo genome editing in the lungs and spleen by disrupting Cas9 activity in the liver. *Nat Biomed Eng* **6**, 157-167 (2022).
550. Rosenblum, D. *et al.* CRISPR-Cas9 genome editing using targeted lipid nanoparticles for cancer therapy. *Sci Adv* **6**(2020).
551. Qiu, M. *et al.* Lipid nanoparticle-mediated codelivery of Cas9 mRNA and single-guide RNA achieves liver-specific in vivo genome editing of Angptl3. *Proc Natl Acad Sci U S A* **118**(2021).
552. Foss, D.V. *et al.* Peptide-mediated delivery of CRISPR enzymes for the efficient editing of primary human lymphocytes. *Nat Biomed Eng* **7**, 647-660 (2023).
553. Herrera-Barrera, M. *et al.* Peptide-guided lipid nanoparticles deliver mRNA to the neural retina of rodents and nonhuman primates. *Sci Adv* **9**, eadd4623 (2023).
554. Kumar, D. *et al.* Tet1 Oxidase Regulates Neuronal Gene Transcription, Active DNA Hydroxy-methylation, Object Location Memory, and Threat Recognition Memory. *Neuroepigenetics* **4**, 12-27 (2015).
555. Gasiunas, G. *et al.* A catalogue of biochemically diverse CRISPR-Cas9 orthologs. *Nat Commun* **11**, 5512 (2020).

556. Sakuma, T., Nishikawa, A., Kume, S., Chayama, K. & Yamamoto, T. Multiplex genome engineering in human cells using all-in-one CRISPR/Cas9 vector system. *Sci Rep* **4**, 5400 (2014).
557. Sakuma, T., Sakamoto, T. & Yamamoto, T. All-in-One CRISPR-Cas9/FokI-dCas9 Vector-Mediated Multiplex Genome Engineering in Cultured Cells. *Methods Mol Biol* **1498**, 41-56 (2017).
558. Klein, C.B. & Costa, M. DNA methylation, heterochromatin and epigenetic carcinogens. *Mutat Res* **386**, 163-80 (1997).
559. Hinz, J.M., Laughery, M.F. & Wyrick, J.J. Nucleosomes Inhibit Cas9 Endonuclease Activity in Vitro. *Biochemistry* **54**, 7063-6 (2015).
560. Isaac, R.S. *et al.* Nucleosome breathing and remodeling constrain CRISPR-Cas9 function. *Elife* **5**(2016).
561. Horlbeck, M.A. *et al.* Nucleosomes impede Cas9 access to DNA in vivo and in vitro. *Elife* **5**(2016).
562. Chen, X. *et al.* Probing the impact of chromatin conformation on genome editing tools. *Nucleic Acids Res* **44**, 6482-92 (2016).
563. Daer, R.M., Cutts, J.P., Brafman, D.A. & Haynes, K.A. The Impact of Chromatin Dynamics on Cas9-Mediated Genome Editing in Human Cells. *ACS Synth Biol* **6**, 428-438 (2017).
564. Jain, S. *et al.* TALEN outperforms Cas9 in editing heterochromatin target sites. *Nat Commun* **12**, 606 (2021).
565. Robertson, K.D. *et al.* The human DNA methyltransferases (DNMTs) 1, 3a and 3b: coordinate mRNA expression in normal tissues and overexpression in tumors. *Nucleic Acids Res* **27**, 2291-8 (1999).
566. Byrd, A.K. & Raney, K.D. Protein displacement by an assembly of helicase molecules aligned along single-stranded DNA. *Nat Struct Mol Biol* **11**, 531-8 (2004).
567. Ikeda, Y., Collins, M.K., Radcliffe, P.A., Mitrophanous, K.A. & Takeuchi, Y. Gene transduction efficiency in cells of different species by HIV and EIAV vectors. *Gene Ther* **9**, 932-8 (2002).
568. Qin, J.Y. *et al.* Systematic comparison of constitutive promoters and the doxycycline-inducible promoter. *PLoS One* **5**, e10611 (2010).
569. Slaymaker, I.M. *et al.* Rationally engineered Cas9 nucleases with improved specificity. *Science* **351**, 84-8 (2016).
570. Kleinstiver, B.P. *et al.* High-fidelity CRISPR-Cas9 nucleases with no detectable genome-wide off-target effects. *Nature* **529**, 490-5 (2016).
571. Lee, J.K. *et al.* Directed evolution of CRISPR-Cas9 to increase its specificity. *Nat Commun* **9**, 3048 (2018).
572. Fu, Y., Sander, J.D., Reyon, D., Cascio, V.M. & Joung, J.K. Improving CRISPR-Cas nuclease specificity using truncated guide RNAs. *Nat Biotechnol* **32**, 279-284 (2014).
573. Hiranniramol, K., Chen, Y., Liu, W. & Wang, X. Generalizable sgRNA design for improved CRISPR/Cas9 editing efficiency. *Bioinformatics* **36**, 2684-2689 (2020).
574. Kolendowski, B. *et al.* Genome-wide analysis reveals a role for TDG in estrogen receptor-mediated enhancer RNA transcription and 3-dimensional reorganization. *Epigenetics Chromatin* **11**, 5 (2018).
575. Ray, S. *et al.* A mechanism for oxidative damage repair at gene regulatory elements. *Nature* **609**, 1038-1047 (2022).
576. Grollman, A.P. & Moriya, M. Mutagenesis by 8-oxoguanine: an enemy within. *Trends Genet* **9**, 246-9 (1993).

577. Fleming, A.M. & Burrows, C.J. DNA modifications walk a fine line between epigenetics and mutagenesis. *Nat Rev Mol Cell Biol* **24**, 449-450 (2023).
578. Perillo, B. *et al.* DNA oxidation as triggered by H3K9me2 demethylation drives estrogen-induced gene expression. *Science* **319**, 202-6 (2008).
579. Fleming, A.M., Ding, Y. & Burrows, C.J. Oxidative DNA damage is epigenetic by regulating gene transcription via base excision repair. *Proc Natl Acad Sci U S A* **114**, 2604-2609 (2017).
580. Pan, L. *et al.* Oxidized Guanine Base Lesions Function in 8-Oxoguanine DNA Glycosylase-1-mediated Epigenetic Regulation of Nuclear Factor kappaB-driven Gene Expression. *J Biol Chem* **291**, 25553-25566 (2016).
581. Cooper, D.N., Mort, M., Stenson, P.D., Ball, E.V. & Chuzhanova, N.A. Methylation-mediated deamination of 5-methylcytosine appears to give rise to mutations causing human inherited disease in CpNpG trinucleotides, as well as in CpG dinucleotides. *Hum Genomics* **4**, 406-10 (2010).
582. Wang, D. *et al.* Active DNA demethylation promotes cell fate specification and the DNA damage response. *Science* **378**, 983-989 (2022).