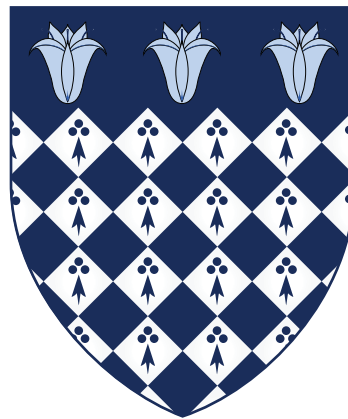


Elucidating the effects of DNA modification on transcription

A thesis submitted for the degree of
Doctor of Philosophy



Olena Yavorska

Magdalen College
University of Oxford

Jointly supervised by
Prof. Skirmantas Kriaucionis and Dr. Benjamin Schuster-Böckler

Michelman 2021

Declaration

This thesis is the result of my own work unless otherwise stated and does not constitute part of any other thesis. The work herein described was carried out at the Ludwig Institute for Cancer Research, under the supervision of Prof. Skirmantas Kriauciūnis and Dr. Benjamin Schuster-Böckler.

Olena Yavorska
Magdalen College
2021

Elucidating the Effects of DNA Modifications on Transcription

Olena Yavorska

Magdalen College

2021

Abstract

DNA modifications play important roles in genome function of vertebrates. 5-methylcytosine (mC) has been extensively studied over the last 50 years as the most common covalent DNA modification, present at roughly 70% of CpGs in the human genome. At promoters it is mostly associated with silencing of genes. An oxidative product of mC, 5-hydroxymethylcytosine (hmC), was found to be common in brain cells in 2007, and studies have attempted to elucidate the function of this stable modification ever since. Although gene silencing through promoter-proximal DNA methylation has been extensively described, the correlation between abundance of mC and hmC in gene bodies and expression is less well understood.

In this study we investigate the relationship between DNA modifications, active transcription, and gene expression. Using a system of global mC and hmC perturbation, we contrast the effects of mC loss ($\text{mC} \rightarrow \text{C}$) and hmC gain ($\text{mC} \rightarrow \text{hmC}$) on different transcription stages: initiation (sense and antisense), pausing, and elongation. The induced changes in transcription are modest but proportional to the changes in DNA modifications, and observations are consistent across multiple analyses.

mC loss is more disruptive to actively transcribing RNA polymerase (RNAP) occupancy and gene expression compared to hmC gain. The effects of hmC gain are modest, but both mC loss and hmC gain result in upregulated initiation (both canonical and spurious), elongation, pause-release, and gene expression. Overall we observe quantitative differences between loss of mC and gain of hmC suggesting that hmC may have a role beyond intermediate of demethylation.

Acknowledgements

This thesis is the culmination of four years of reading, learning, debugging, and plotting. However, as the pandemic highlighted, having the right network of people around you is what keeps you progressing despite the inevitable intermittent frustration.

First and foremost, I would like to thank my two supervisors, Prof. Skirmantas Kriauciūnas and Dr. Benjamin Schuster-Böckler, for guiding and encouraging me throughout the last four years. I am grateful that you didn't mould me, but rather allowed me to become my own researcher: relatively sceptical, too enthusiastic about figures, passionate about parallelisation of cluster jobs, and with slight tendencies to jump down rabbit holes. Thank you for listening to and discussing my random scientific musings, even if they weren't always relevant (most of the time they weren't really relevant...). I've always felt like I had mentors rather than bosses, and so I have learnt an immense amount during my DPhil, not only about science but also the qualities that make a great teacher. I am genuinely so grateful that I got the opportunity to work under your guidance.

Next, I would like to thank my examiners, Prof. Rob Klose and Dr. Gabriele Schweikert, for taking the time to read this thesis in detail and for their suggestions and advice.

This thesis would contain very little data without Dr. Hiromi Tagoh. Hiromi, thank you so much for all your scrupulous work on this project. Analysis is always easier when the experiments went well and data has been generated carefully, and you made my job much easier than it could have been. Thank you for your input and feedback throughout and for always answering my silly questions. I feel very lucky to have worked with you on this project.

I would like to thank my two labs for being incredibly welcoming, for their feedback, for the copious baked goods over the years, and for the laughs. In particular, I would like to thank Dr. Paolo Spingardi and Dr. Paulina Siejka-Zielinska for generating some of the data presented in this thesis and Dr. Pijus Brazauskas for designing the inducible cell lines used in this project.

I would also like to thank my previous supervisors. Dr. Wolfgang Huber and the

Huber group, thank you for introducing me to this wonderful world of biology, for taking me in as a mathematician who had very little knowledge about genetics and programming, and basically defining my future career through an immensely enjoyable four month internship at EMBL Heidelberg. Another special thanks to Dr. Stephen Burgess for taking a chance on me through a reddit-application and allowing me to stretch my programming skills into something tangible and useful to the Mendelian randomisation community.

Magdalen College, thank you for being my home throughout this process, with a special thanks to Andy Higgs for the numerous candid chats in the Porters' Lodge. I have met some incredible, lifelong friends through being part of your community.

(Almost Dr.) Ryan Hoyle, you have been a beacon of optimism since I met you in Freshers' week. I am continually mesmerised by your openness to new experiences, to helping people, to spending literal hours discussing how all the information from a DESeq2 results object could be included in one plot. Thank you for the walks, the movie nights, the salsa, and the lengthy chats about everything and anything.

(Almost Dr.) Jason Weir, you have been a wonderful friend since we first discovered that we were housemates in Michelmas 2017. Thank you for teaching me about Wagner, for your passion for specific broccoli cooking directions, and for often being my voice of reason over the last four years.

Nathan Davies, you are truly one of a kind. Thank you for the 7am walks around Addison's, for your kindness, and for opening up the weird world of philosophy to me.

Dr. Dean Frawley, thank you for our long walks, our Stranger Things and (failed) Lord of the Rings marathons, but mostly for your unique ability to encourage self-reflection. Liam McDonnell, Kevin Zoka, and John Granahan, thank you for periodically reminding me that "it'll be grand sure" throughout a worldwide pandemic, often that's all that's needed. Hannah Maguire and Sarah Flannery, thank you for the Eurovision-Umi-facemask ridiculousness that has defined our friendship for over fifteen years now.

Michael Hinge, thank you for the daily laughing fits, the encouragement, the slipper-wearing, and for being one of the most incredible people I've ever met. Even with the

world still slightly in crisis, these last few months have genuinely been some of my happiest thanks to you. I love you.

Mam and Dad, thank you for giving up so much to allow me to get to this point. You prioritised our education over everything as if it was the simplest decision in the world. Mam, since learning off reams of dialogue with you for playschool, your patience, encouragement, and diligence instilled a life-long work ethic in me to always power through and not avoid things that may be difficult. Dad, your problem-solving approach to academia and life in general (sometimes at the expense of DVD players...) ensured that I always tried to think outside the box. You both define who I have become as a researcher, and as a person, and without both of you, none of this would have been possible.

This thesis is dedicated to my grandparents, Prof. Petro Kalenyuk and Hanna Kalenyuk, Assoc. Prof. Lidia Yavorska and Assoc. Prof. Igor Yavorsky.

Contents

1	Introduction	1
1.1	Mechanisms that guide the phenotype of a cell	1
1.2	DNA Modifications	3
1.2.1	Prevalence and role	3
1.2.1.1	CpG Islands	4
1.2.2	Regulation	4
1.2.2.1	Where do they come from?	5
1.2.2.2	Active Demethylation	6
1.2.2.3	5-hydroxymethylcytosine (hmC)	7
1.2.2.4	5-carboxymethyl cytosine (caC) and 5-formylcytosine (fC)	9
1.2.3	Techniques to detect DNA modifications	9
1.3	Transcription	12
1.3.1	RNA Polymerases	12
1.3.2	Stages of transcription	12
1.3.2.1	Recruitment and Initiation	12
1.3.2.2	Pausing	14
1.3.2.3	Elongation and Termination	15
1.3.2.4	Other mechanisms	17
1.3.3	Techniques to measure engaged RNA polymerase occupancy	18
1.4	Influence of DNA modifications on transcription	22
1.4.1	Aim of thesis	23
2	Methods	25
2.1	Annotation	25
2.1.1	Genome annotation	25
2.1.1.1	Metagenes	25
2.1.2	CpGI annotation	25
2.1.3	Enhancer annotation	26
2.1.3.1	FANTOM5	26
2.1.3.2	EnhancerAtlas v2.0	26

2.1.3.3	Enhancer classification	27
2.1.3.4	Enhancer annotation comparison	27
2.2	DNA Modifications	27
2.2.1	Pre-processing of TAPS and TAPS- β data	27
2.2.2	Modification level simulation	29
2.2.3	Metagenes	31
2.2.4	Genome Browser	31
2.2.5	Differential methylation analysis	31
2.3	mRNA	31
2.3.1	Pre-processing of mRNA data	31
2.3.2	Differential expression analysis of mRNA count data	32
2.4	Engaged polymerase occupancy	33
2.4.1	QC analysis and deduplication	33
2.4.2	The reference genome	34
2.4.3	Processing the .bam files	35
2.4.3.1	Generation of signal tracks	35
2.4.3.2	Downsampling PRO-Cap and PRO-Seq data	35
2.4.3.3	Metaplot generation	35
2.4.4	Identifying regions of polymerase initiation	35
2.4.4.1	The HMM	35
2.4.4.1.1	Counts, observed variables, and blocking	36
2.4.4.1.2	Transition matrices, peak calling, and trimming	36
2.4.5	Linking initiation regions to features	38
2.4.6	Identifying regions of polymerase elongation	38
2.4.7	Selecting major initiation regions for each gene	38
2.4.8	Assigning probable elongation regions	38
2.4.9	Transcription statistics	38
2.4.10	Uniformity index to evaluate stalling	39
2.4.11	Termination index	39
2.4.12	Differential RNAP initiation and occupancy analysis	39

3	Induction of unbiased genome-wide DNA modification changes	41
3.1	Introduction	41
3.2	Results	42
3.2.1	Experimental Conditions	42
3.2.1.1	TET1-CD upregulation via Dox induction	42
3.2.1.2	Demethylation with 5-aza-CdR	43
3.2.1.3	Isolating the effects of 5hmC induction from those of demethylation	44
3.2.2	Detection of base-resolution DNA modification changes in HEK293 cells	45
3.2.2.1	Quality control	45
3.2.2.1.1	Conversion rate of spike-ins	45
3.2.2.1.2	Coverage	54
3.2.2.1.2.1	Copy Number Analysis	55
3.2.3	Genome-wide base-level analysis of induced modification changes	56
3.2.3.1	Calculating hmC level and changes in mC	57
3.2.3.2	The effects of incomplete conversion	61
3.2.4	Genome-wide analysis of trends in induced modification changes	64
3.2.5	Analysis of DNA modifications in functional regions	65
3.2.5.1	DNA modification changes in genes	65
3.2.5.2	DNA modification changes in CpG islands	69
3.3	Discussion	77
4	The effect of genome-wide DNA modification changes on occupancy of ac- tively engaged RNA polymerase	80
4.1	Introduction	80
4.2	Results	80
4.2.1	Experimental set-up	80
4.2.2	Pre-processing and deduplication	84
4.2.3	Sample normalisation and spike-ins	86
4.2.4	Overview of initiation and occupancy of active RNAP in HEK293 cells	87
4.2.5	Are RNAP initiation and occupancy metagene profiles affected by DNA modification changes?	90

4.2.6	Annotation Agnostic Analysis	90
4.2.7	Binned initiation	91
4.2.8	Data-driven initiation	91
4.2.8.1	HMM model	91
4.2.8.1.1	HMM design	93
4.2.8.2	Trimming	95
4.2.8.3	Stringency and variation analysis	95
4.2.8.4	Pausing	96
4.2.8.5	Run-through	99
4.2.8.6	Elongation	103
4.2.9	Supervised analysis of RNAP initiation and elongation	106
4.2.10	TSS identification	107
4.2.10.1	Transcript annotation analysis	108
4.2.10.2	Defining initiation regions	110
4.2.10.3	Definition of elongation regions	120
4.2.11	Effects of DNA modification	122
4.2.11.1	Treatment-induced changes in initiation	122
4.2.11.1.1	Is PRO-Cap representative of initiation frequency?	122
4.2.11.1.2	Treatment-induced changes in PRO-Cap signal	123
4.2.11.2	Analysis of termination, pausing, and uniformity	133
4.2.11.2.1	Termination	133
4.2.11.2.2	Pausing regions and pausing index (PI)	135
4.2.11.2.3	Stalling and uniformity index (UI)	136
4.2.11.3	Treatment-induced changes in elongation	138
4.3	Discussion	143
5	Relationship between engaged RNA polymerase occupancy and mRNA abundance after genome-wide DNA modification changes	148
5.1	Introduction	148
5.2	Results: mRNA analysis	149
5.2.1	mRNA abundance estimation	149

5.2.2	Experimental design : Mitigating DNA methylation independent treatment effects	150
5.2.3	Count normalisation	152
5.2.4	Differential expression analysis	152
5.2.5	Comparing gene expression changes after +Dox and +5-aza-CdR . . .	154
5.2.5.1	Focus on anti-sense lncRNA	158
5.3	Results: Combining mRNA abundance and DNA modification data . . .	161
5.3.1	Relationship between baseline expression and mC and hmC levels . . .	161
5.3.2	Do genes with changes in expression show parallel changes in mC and hmC abundance?	163
5.4	Results: Combining mRNA abundance and RNA polymerase occupancy data	168
5.4.1	Baseline PRO-Seq as a predictor of baseline mRNA	168
5.4.2	Are changes in PRO-Seq predictive of changes in mRNA?	170
5.5	Discussion	177
6	Discussion of future directions	180
6.1	Changes to experimental design	180
6.2	Alternative and additional experiments	182
6.3	Related questions	182
7	Conclusion	183
	References	184

1 Introduction

1.1 Mechanisms that guide the phenotype of a cell

Human DNA is long and complex, but changes in the nucleotide sequence alone cannot explain the biological variability between cells [160]. The same DNA can produce different phenotypes of cells by selective expression of genes, which are activated and deactivated through highly dynamic regulation (Figure 1).

In order to synthesise a protein, the relevant gene(s) must be active and accessible for transcription. Chromatin is made up of structures called nucleosomes, comprised of DNA tightly wrapped around an octamer of histone proteins (two of each: H3, H4, H2A, and H2B). The histones have N-terminus tails which can be modified, and one function of these modifications is the regulation of how open the DNA is and thus which subset of genes can be active.

Regulation of gene expression is facilitated through a subset of transcription factors (TFs) that can cooperate with chromatin remodellers which alter histone modifications. Of the TFs that recruit chromatin modifiers¹, most aid transcription through exposing promoters, while others play a repressive role.

TF binding is highly targeted, thus the presence of DNA methylation in regulatory regions can affect transcription. Some TFs can directly bind to regulatory regions on the DNA to modulate transcription. While methyl-CpG-binding proteins (MBPs) favour binding at methylated regions, most proteins preferentially bind unmethylated DNA. Therefore DNA modifications can affect transcription through disrupting TF binding.

TFs facilitate recruitment of RNA polymerase and the transcriptional machinery through binding at enhancers and promoters, thus sensitivity to DNA modifications can result in reduced binding, and therefore gene repression. In this project, we focus on the relationship between DNA modifications and active transcription. In other words we investigate how DNA modifications affect gene expression by perturbing DNA modifications and measuring RNAP occupancy.

¹Note that TFs have many roles independent of this.

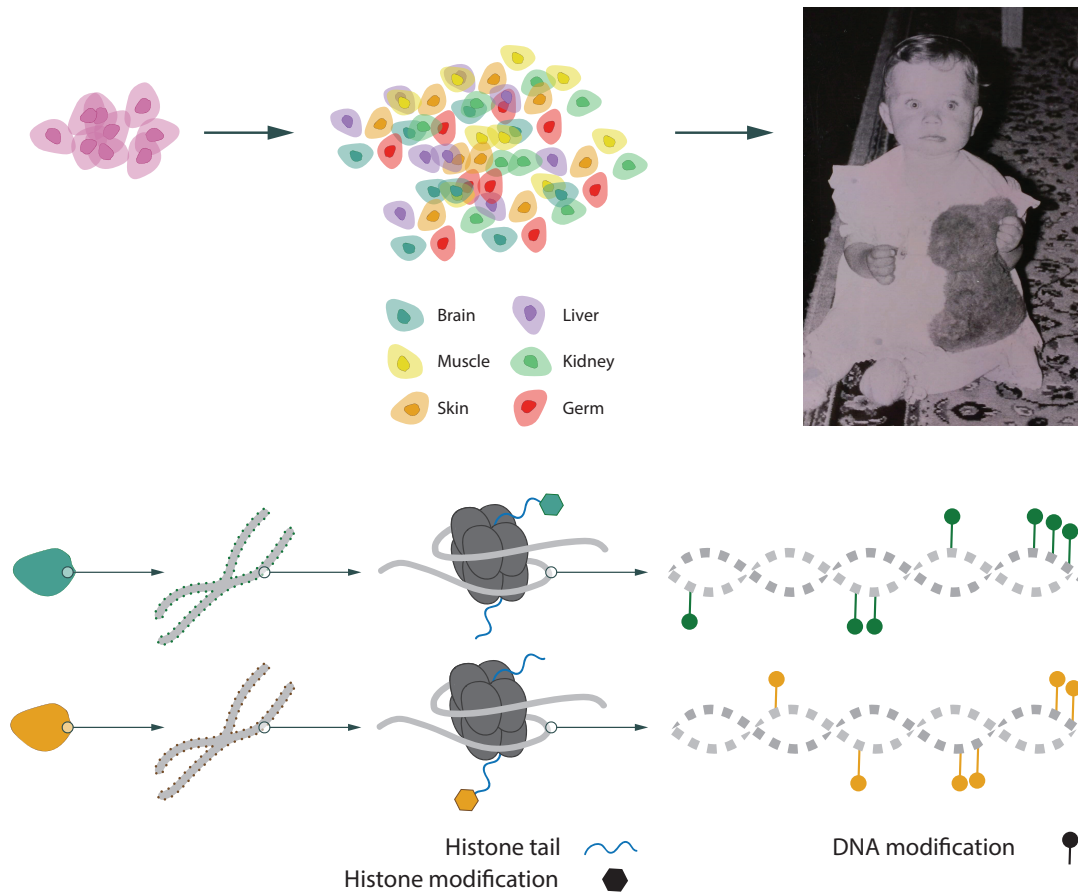


Figure 1: **Epigenetics and phenotype.** To guide a cell from its progenitor state to a differentiated mature cell with defined function in an organism, the cell must decide which proteins are synthesised through regulation of gene expression. One method of regulation involves modifications “upon” (or *epi*) the DNA. On the chromatin level, by inducing changes to chromatin compaction through histone tail modification one can change which regions are accessible to transcription factors in a cell-specific manner. Similarly, the DNA bases can be modified directly in a heritable way without changes to the underlying DNA sequence allowing for dynamic regulation at a finer level.

1.2 DNA Modifications

1.2.1 Prevalence and role

The most prevalent DNA modification (DNAm_{od}), 5-methylcytosine (mC) was first described by G.R. Wyatt in 1950 [263] and is found in all known vertebrates and land plants [269]. This modification is stable in the genome, being faithfully copied after each replication cycle, and is thought to be involved in regulation of gene expression [7, 79, 110, 229] and consequently cell identity stabilisation [211], transposable element silencing [26, 86, 115, 148] and X chromosome inactivation [96, 199, 217, 235], imprinting [58, 134, 242], and chromatin structure through nucleosome stability [35, 37, 189].

The prevalence and types of DNAm_{ods} vary across species. Although CpG dinucleotides are under-represented in the genome, mammalian DNA methylation (DNAm_{eth}) is most prevalent in this context with 70-80% of CpGs being symmetrically methylated (mostly in repetitive regions)[57]. In contrast, methylation in plants often occurs outside CpG contexts (reviewed in [92]). While there is evidence of non-CpG methylation in many tissue types across mammalian genomes, it is significantly less frequent² (reviewed in [182]).

DNAm_{eth} has been extensively studied within the scope of differentiation. In particular, multiple studies examining the dynamic changes in methylation through a hematopoietic differentiation have found cell-type specific methylomes (reviewed in Ko et al. [119]). Hodges et al. [95] found that cell-type specific hypomethylation often marked lineage-specific genes and, through the course of differentiation, intermediate methylation states tended to a uniform (hypo- or hyper-methylated) state in the mature cell. Bock et al. [22] showed that methylation differences between cell types were modest but *locus-specific* and often overlapped with transcription factor (TF) binding sites, suggesting a role in preventing aberrant TF binding. Based on this, cells could be classified based on their methylome.

DNAm_{eth} is usually associated with repression. In their model, Sharifi-Zarchi et al. [216] showed that if DNAm_{eth} is present in a region then it's inactive, whereas if

²This is nevertheless an area of active research, in particular for mitochondrial DNA which shows extensive methylation in non-CpG contexts [183].

H3K4me3 or H3K4me1 are abundant then the region is a promoter or enhancer, respectively³. While hypermethylation suggests inactivity, hypomethylation is not always marker of an active region. Hypomethylated regions are indeed H3K4me3 enriched, but not always [216]. Indeed, Hodges et al. [95] found that although most hypomethylated promoters are common across different cell types, this hypomethylation does not necessarily define that the related genes are active.

1.2.1.1 CpG Islands

The methylation landscape in humans depends largely on CpG density. In their investigation of the effect of topology on methylation, Lövkvist et al. [155] showed that CpG clusters are rarely found in an intermediate methylation state, in other words a CpG cluster is either hypo- or hyper-methylated. They also noted that while methylated CpGs can affect the methylation of an extended local region, unmethylated CpGs only affect the demethylation of nearby CpGs.

Stretches of DNA with a high CpG density are called CpG islands (CpGIs), and remain largely unmethylated⁴. Lövkvist et al. [155]’s findings suggest that CpGIs are short and dense to ensure efficient demethylation, and since they overlap with ~70% of human gene promoters a role for transcriptional regulation also seems likely [232, 207].

Indeed, hypermethylation of CpGIs, or CpG island methylator phenotype (CIMP), was first described for primary colorectal cancers in 1999 [240] and has since been found in both normal and cancerous tissues (reviewed in [102, 166]). Although most CpG islands overlap with promoters, gene-body CpGI hypermethylation has also been linked to tumorigenesis [6]. In fact there exists a paradox whereby although promoter CpGI methylation is repressive, gene-body CpGI methylation is associated with transcription [109].

1.2.2 Regulation

³H3K4me3 is associated with active promoters, and H3K4me1 is an enhancer marker.

⁴Technically defined as being 200bp or longer with GC content exceeding 50% and a CpG observed over expected ratio of over 0.6 [74].

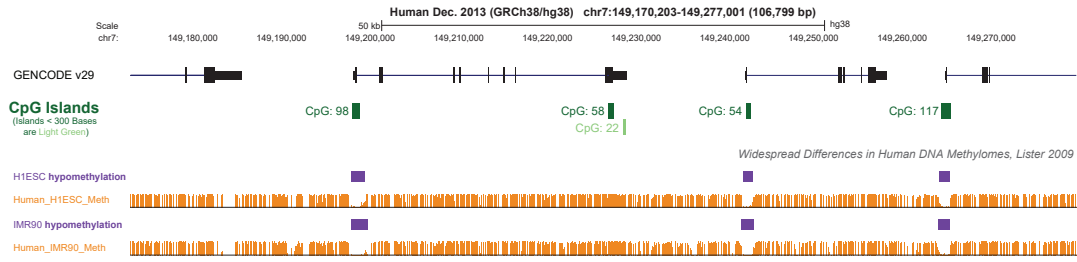


Figure 2: **Example of CpGI hypomethylation.**

1.2.2.1 Where do they come from?

Mammals have three enzymes which are responsible for maintenance of DNAmeth, known as DNA methyltransferases (DNMTs). In 1999, Okano et al. [177] described two variants, Dnmt3a and Dnmt3b, which were responsible for *de novo* methylation of mouse cells. These enzymes preferentially localise near CpG dense regions but are excluded from active promoters and enhancers, with Dnmt3b showing higher binding at gene bodies of active genes [16]. The maintenance of mC is carried out by DNMT1 which has a preference for hemimethylated CpGs (but has also shown some evidence of *de novo* methylation) [267].

The evolutionary importance of DNAmeth was recently studied in yeast, where Dnmt5 is a highly specific maintenance enzyme but the *de novo* DNMT was lost 50-150 million years ago. Catania et al. [31] showed that despite the loss, the species shows low methylation levels which have been propagated and suggested an evolutionary, natural selection explanation for the persistence of DNAmeth in the genome. However, in most cases absence of these enzymes is lethal to cells [133, 177], and silencing can have very disparate consequences depending on cell-type and organism.

All three enzymes are essential, and mice lacking Dnmt1, Dnmt3a, or Dnmt3b are embryonic lethal [133, 177]. Without DNMTs, methylation cannot be effectively copied and the genome becomes progressively demethylated after each cell division. Even insufficient quantities of Dnmt1 can lead to significant problems, e.g. mice carrying a modified Dnmt1 allele which resulted in 10% of the normal mC developed aggressive T cell lymphomas at 4-8 months [75] and knock-out in specific cell types has a similarly

poor prognosis. In their study of gastrointestinal development, Jorgensen et al. [113] used smooth-muscle Dnmt1-KO mice and found that the pups were smaller than their litter mates, had massive intestinal dilation, and died around postnatal day 21. In contrast, bees with silenced Dnmt3 emerge preferentially as queens rather than workers [124] and Dnmt3 knockdown in locusts results in changes in phenotypic plasticity (affecting swarming) [98].

In some cases DNMTs cooperate and compensate for one another. Rhee et al. [197] showed that while deletion of DNMT1 reduced global methylation by less than 3% in a colorectal cancer cell line, deletion of both DNMT1 and DNMT3b reduced methylation by over 95%. In their study of postmitotic neurons, Feng et al. [64] found that only double knockout (Dnmt1 and Dnmt3a) mice showed changes in plasticity and learning, with smaller neuronal phenotypes and changes in gene expression and methylation compared to wild type. Combined with a previous study of Dnmt1 deletion in postmitotic neurons which failed to show any obvious phenotypic changes [63], Dnmt3 appears to compensate for Dnmt1 in these cells too.

1.2.2.2 Active Demethylation

Demethylation can occur passively through replication if activity of DNMT enzymes is compromised, but, an active way to remove methyl groups is necessary for more targeted regulation. The enzymes responsible for active demethylation remained elusive until Tahiliani et al. [234] identified the ten-eleven translocation 1 (TET1) enzyme as a candidate demethylase. They rationalised that since the TET1 homologs in *Trypanosoma* oxidised 5mT, the enzyme may serve a similar oxidative function in humans, and were correct. Ito et al. [103] then showed that TET1, TET2, and TET3 are also the enzymes responsible for active demethylation in mice.

The TET enzymes combined with base excision repair (BER) form an active pathway that can convert mC back to cytosine [128, 234, 45, 253] (Figure 3). They oxidise mC to 5-hydroxymethylcytosine (hmC), then to 5-formylcytosine (fC), and lastly to 5-carboxylcytosine (caC) [122, 90, 104]. Since most DNAm are found in a CpG context, it is not surprising that DeNizio et al. [47] found TET2 mC oxidation is more efficient

at CpG compared to CpH contexts (the preference is reduced at hmC). The oxidative products of mC prevent DNMT1 from copying the methylation onto the daughter strand due to reduced catalytic activity, suggesting a coordinated demethylation through both the active and passive pathway [214].

Thymine DNA glycosylase (TDG) acts on T:G, U:G, and hmU:G mismatches, but has also been shown to convert fC and caC to an abasic site with little preference for context [90, 161, 47]. The BER pathway then excises the abasic site and a polymerase replaces the base by an unmodified cytosine, completing the active demethylation process. TDG is essential in completing the conversion of mC to cytosine (reviewed in [45]). The enzyme is required for mouse development, with *TDG*^{-/-} mice exhibiting an embryonic lethal phenotype [41, 42].

Experiments attempting to differentiate cells with TET deficiencies or perturbations have shown that these enzymes are essential for lineage specification in blood [2, 157], neurons (reviewed in [204]), and ESCs [139]. TET1 mostly localises near CpG-rich promoters and within genes [257], resulting in reduced methylation both through its hydroxylase activity but also by physically obstructing DNMT from binding. The relationship between TET-induced hypomethylation and gene expression is not always positively correlated and repressive roles of Tet1 have been proposed by multiple groups. Wu et al. [262] showed that Tet1 promotes PRC2 recruitment in mouse ESCs, facilitating H3K27 methylation, and [257] showed colocalisation with the SIN3A co-repressor complex which mediates histone deacetylation.

It is worth noting that more recent studies have found that mC can be oxidised by the DNA repair enzymes ALKBH2, ALKBH3, and AlkB in vitro [19].

1.2.2.3 5-hydroxymethylcytosine (hmC)

Originally discovered in simpler organisms in 1952 by Wyatt and Cohen [264], interest in hmC was revived in 2009 after being found in high abundance in Purkinje neurons (~40% of mC) by Kriaucionis and Heintz [122] and in mouse ESC (~10% of mC) [234]. Although total abundance is highly variable across cell types [77], it is most prevalent in the brain, and levels are lower in cancer than in healthy tissue [9]. Loss of hmC has

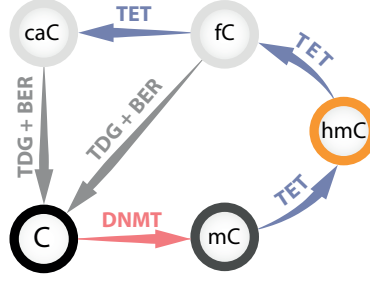


Figure 3: **Active demethylation pathway.** Unmodified cytosine (uC) is methylated by DNMTs. The TET enzymes oxidise mC to hmC, fC, and caC. The latter, fC and caC, are unstable and removed by TDG as part of the BER pathway.

been associated with more aggressive cancers [272, 141]. Recently Wang et al. [251] showed global reduction of hmC in lung cancers, with concurrent increases in hmC at some CpG islands and gene promoters, and Lian et al. [140] defined loss of hmC as an “epigenetic hallmark of melanoma”.

hmC is most abundant in gene bodies, with some enrichment in proximal upstream and downstream regions relative to transcription start sites (TSSs), transcription end sites (TESs) and distal regions [222]. Although TSSs are generally free of mC, hmC appears to be enriched in TSSs, in particular those that were CpG-rich and had H3K27me3 and H3K4me3 marks [181, 257]. Although high concentration of hmC at promoters is still repressive in general [203, 181], Ficz et al. [66] showed an association between hmC at TSSs and increased transcription.

There is more evidence of gene body hmC being a marker of high transcription [144, 165, 257, 262, 241]. During construction of their hmC tissue map, Cui et al. [44] found that tissue-specific gene bodies and enhancers have the most hmC. Enrichment in enhancers (marked by H3K4me1 and H3K27ac) was also noted by Stroud et al. [231].

Similarly to mC, the levels of hmC can change during differentiation [243], but unlike its oxidative products it is a stable modification [83, 193, 233]. The stability of hmC is exemplified in recent studies which have successfully used hmC maps from circulating cell-free DNA (cfDNA) to classify cancers, suggesting its potential in diagnostics and personalised medicine [138, 33, 265].

Other roles for hmC, independent of transcription regulation, have been proposed. The relationship between hmC and DNA repair is another active area of research, reviewed in [116]. Kafer et al. [114] suggested a role in genome integrity, namely that hmC is a marker for DNA damage, however deconvolution of the effect of hmC and TET was not clear. In 2019, Mohni et al. [167] discovered a hmC binding ESC-specific (HMCES) reader of abasic sites without which AP site repair is delayed and DNA damage accumulates, and Shukla et al. [219] noted class switch recombination defects in HMCES deficient mice. In an effort to explain the link between hmC and recombination, Robertson et al. [202] described a model whereby hmC is a preferred substrate for endonuclease G, which catalyses the formation of double stranded DNA breaks, promoting recombination.

1.2.2.4 5-carboxymethyl cytosine (caC) and 5-formylcytosine (fC)

The abundances of fC and caC are very low, Globisch et al. [77] estimated fC and caC abundance at roughly 1.4% and 0.3% of hmC, respectively. This is both a consequence of being difficult to oxidise by TET [99] and being excision targets of TDG [90, 161, 186, 271].

As fC and caC are marked for removal by the BER machinery, mapping these modifications can provide an insight into which regions are the most actively demethylated. Song et al. [223] showed that fC appears to be enriched in mESCs at active and poised enhancers (with a preference for poised⁵).

1.2.3 Techniques to detect DNA modifications

Many techniques have been developed with the goal of studying modified cytosines in the genome. By the late 1980s, simple measurements of the amount of mC in a given genome could be calculated and some methylation sensitive restriction enzymes had also been identified [85]. Fast forward, and methylation arrays were used to quantify methylation in specific regions of the genome, or alternatively one could pull down the regions of interest with methylated DNA immunoprecipitation (MeDIP) [255]. However,

⁵Marked by H3K4me1 but not H3K27ac.

accurate methylation mapping at base resolution has been synonymous with bisulfite sequencing (BS) for a long time.

In 1970, two groups independently discovered bisulfite-mediated deamination of cytosine ([215],[88],[85]). Hayatsu et al. [89] observed mC can be similarly deaminated, but this process takes much longer [87], providing a way to distinguish cytosine from mC. During sequencing, bisulfite-treated cytosines are read as thymines whereas mCs remain as Cs allowing for bioinformatic separation of C and mC.

In 2009, Kriaucionis and Heintz [122] were surprised to find relatively high levels of hmdC in Purkinje neurons (0.6% of total nucleotides) and a general enrichment of this DNA modification in brain. The authors used a combination of chromatography and mass spectrometry to quantify the abundance of the modification, but techniques for more targeted mapping of hmC soon emerged.

Although mC (being the most abundant DNA modification) was usually the main focus of BS analyses⁶, hmC behaves like mC in that it's not read as T (fC and caC behave like cytosine). Therefore BS could be used to detect mC/hmC positions, but not mC or hmC alone. In 2012, Booth et al. [25] proposed a derivative of BS, called oxidative BS (ox-BS), where the hmC is first specifically oxidised to fC by K₂Cr₂O₇. During the subsequent bisulfite treatment, only mC's remain as cytosines when sequenced, therefore one can use a combination of BS and ox-BS to discriminate between mC and hmC. To detect hmC only, Tet-assisted bisulfite sequencing (TAB-Seq) [268] is a cheaper alternative to ox-BS. In TAB-Seq, mouse TET1 enzymes are used to oxidise any modified cytosines to caC, with the exception of hmC which is first protected by β GT.

In this project, we wanted to map the positions of both mC and hmC, but BS has some drawbacks. It is relatively harsh on the DNA, and therefore assumes large volumes of material (this is not always practical for small patient samples or cell free DNA). More importantly the technique largely reduces a 4-base genome to a 3-base genome for the purposes of mapping by converting non-modified cytosines to thymines, resulting in dips in coverage and reads that cannot be uniquely mapped.

⁶Recalling the relative abundances of the oxidative derivatives of mC from Section 1.2.2, the contribution of hmC to the total number of modified bases is low for most cell types.

In 2019, Liu et al. [150] published a new, bisulfide-free alternative to BS, called TET-assisted pyridine borane sequencing (TAPS) which addressed many of the problems with BS. This method converts mC and its oxidative derivatives, which is less destructive, cheaper, and leaves the majority of the genome unchanged (Figure 4). By first oxidising modified bases (mC and hmC) using TET enzymes, then treating with pyridine borane, one reduces all oxidative derivatives of mC into DHU which is converted to T during the polymerase chain reaction (PCR) step. The authors showed that pyridine borane can be used to convert fC and caC to DHU with 97.6% and 98.1% conversion rates, respectively.

To separate mC and hmC, Liu et al. [150] combined the theory behind ox-BS and TAB-Seq. TAPS- β is a derivative of TAPS which uses β GT to protect hmC from oxidation by TET, leaving mC, caC, and fC as the only converted bases. Similarly to the BS/ox-BS analyses, mC and hmC can be bioinformatically separated by combining TAPS and TAPS- β results (Figure 4).

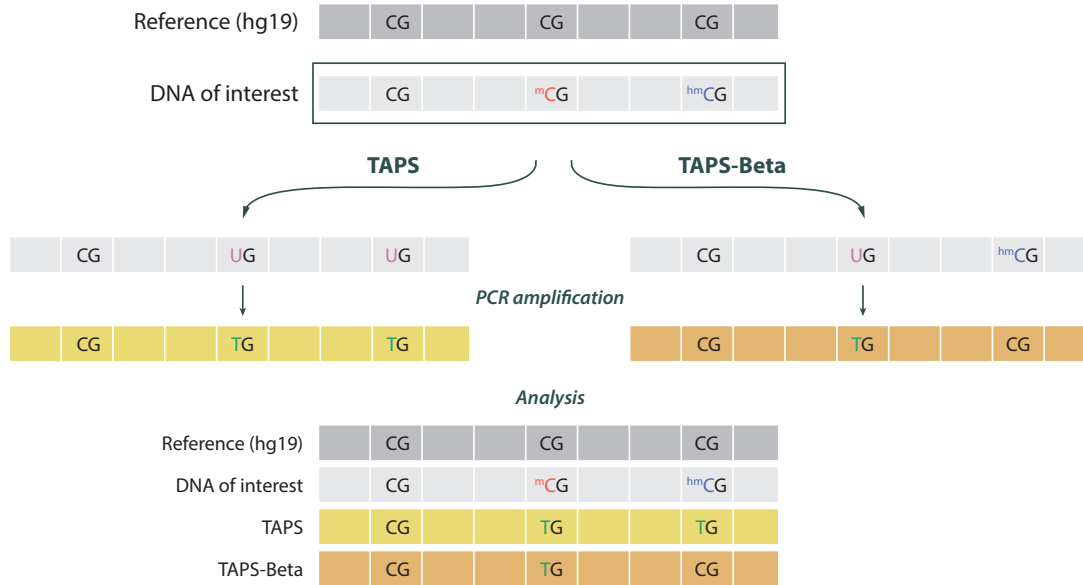


Figure 4: **TAPS and TAPS- β explained.** Schematic showing the expected conversion of cytosines in TAPS and TAPS- β . When comparing to the reference sequence, we expect mC and hmC positions to be mapped as Ts in TAPS but only see Ts at mC in TAPS- β .

1.3 Transcription

1.3.1 RNA Polymerases

There are three RNA polymerases (RNAPs) in metazoans, RNAPI, RNAPII, and RNAPIII, each with distinct roles.

RNAPI and RNAPIII are responsible for transcription of non-coding RNAs, notably ribosomal RNA (rRNA) and transfer RNA (tRNA), respectively. The promoters of these transcripts differ significantly from those of genes transcribed by RNAPII as they are largely uni-directional, and are more “bendable” which is compensated for by the structure of the complex [61].

RNAPII is the most active, and synthesises messenger RNA (mRNA) as well as long non-coding RNA (lncRNA), microRNA (miRNA), and small nuclear RNA (snRNA). After initiation, the nascent strand is usually processed co-transcriptionally (reviewed in [168]) but only a small proportion of RNAPII continue to productive elongation [228]. When modelling RNAPII kinetics, Steurer et al. [228] found that most (~93%) of RNAPII is chromatin-bound with 10% initiating, 23% in a paused state, and the other 60% in the process of elongation.

1.3.2 Stages of transcription

Transcription proceeds in stages (Figure 5), with potential of regulation at each one, resulting in delicate tuning of the timing and rate of mRNA production.

1.3.2.1 Recruitment and Initiation

First, transcription activators bind to enhancers and recruit co-activator complexes which can open up the chromatin. This space is needed to facilitate assembly of the preinitiation complex (PIC), which is comprised of general transcription factors (GTFs) that bind to the promoter. The PIC is stabilised by the Mediator, forming a PIC-Mediator complex [121], which recruits RNAPII to the promoter. For initiation to take place, a subunit of RNAPII is phosphorylated by CDK7 allowing it to escape the

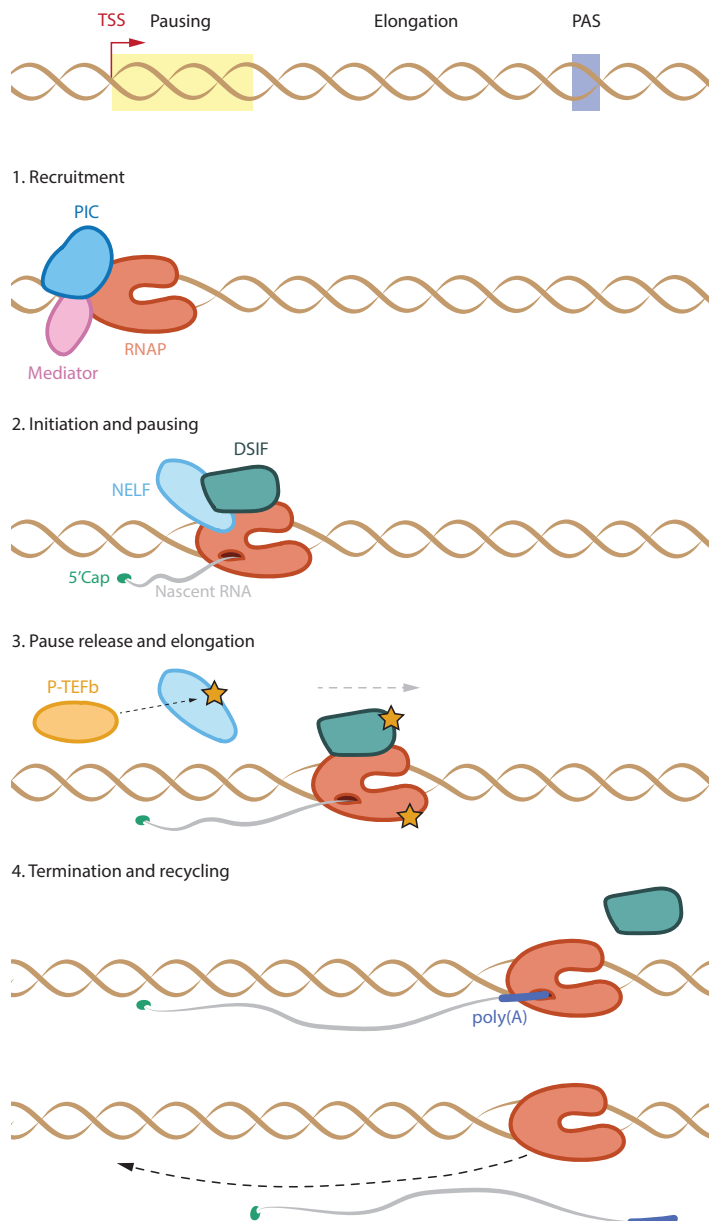


Figure 5: **Stages of transcription.**

1. The pre-initiation complex (PIC) is formed from general transcription factors (GTFs), and together with the Mediator complex recruits RNAP to the promoter. 2. After initiation, the 5' end of the nascent RNA is capped, but RNAP doesn't get very far before pausing factors slow down its rate and cause promoter-proximal pausing. DRB-sensitivity inducing factor (DSIF) and negative elongation factor (NELF) are responsible for stabilising paused RNAP. 3. To release RNAP into productive elongation, positive transcription elongation factor b (P-TEFb) phosphorylates DSIF, NELF, and the C-terminal domain (CTD) of RNAP. This results in NELF dissociating from the DSIF-NELF complex and allows RNAP to continue elongating. 4. Once the polyadenylation site (PAS) is reached, the RNAP pauses again allowing for efficient termination before dissociating.

promoter⁷.

The above is not continuous. A gene can fluctuate between active and inactive states, resulting in discontinuous bursts of transcription whose frequency is regulated by TFs (activators and repressors) and enhancers [127]. The size of the burst is affected by the presence of core promoter elements such as the TATA box and initiator element (Inr)⁸, other features like TA content at 25-30bp upstream of the promoter have also been associated with the strongest promoters [156]. Active TSSs are largely found in clusters of divergent pairs [238] and genes with promoter-promoter interactions are often transcribed cooperatively, with interacting promoters influencing each other Li et al. [135].

RNAPII is also responsible for enhancer RNA (eRNA) transcription. In their model, Bartman et al. [14] showed that the burst (which includes enhancer transcription) actually precedes PIC formation and RNAPII recruitment. In contrast to genes, bursts of eRNA synthesis are often followed by a return to base levels, suggesting that once the promoter has been activated the enhancer is no longer needed Arner et al. [8].

1.3.2.2 Pausing

The transition from initiation to pausing involves the binding of pausing factors DRB-sensitivity inducing factor (DSIF) and negative elongation factor (NELF) which slow down the rate of RNAPII. Promoter-proximal pausing is a common attribute of RNAPI, II, and III, but the distributions of pausing times are different between the polymerases. The shortest pauses are associated with RNAPIII, whereas RNAPII has the highest median pausing time [270].

After recruitment and initiation, only 12.5% of initiating RNAPII continue to promoter-proximal pausing, which occurs 20-80bp⁹ downstream of the initiation site in most promoters but the exact position and proportion of affected promoters differs between studies [228, 38, 126, 238, 39, 80]. In their study, Gressel et al. [80] found that RNAPII

⁷Fujiwara et al. [71] showed that most of the GTFs of the PIC remain associated with RNAPII as far as 27bp downstream of the TSS, but had dissociated by 49bp.

⁸The most transcriptionally active gene promoters contain both.

⁹Pausing closer to the TSS is associated with more enhancer-like transcription, with little elongation [238].

pausing generally occurred at 84bp downstream of the TSS, with pausing sites enriched for GC. Similarly, Watts et al. [252] noted an over-representation of a C-rich 9-mer motif in pause sites¹⁰. The relationship with sequence composition explains why pausing is relatively consistent between cell types, with 19% of genes showing pausing maxima within 3bp [252].

Based on an analysis of 3200 genes, [111] calculated the average half-life of paused RNAP at 7 minutes, suggesting that the paused complex is relatively stable, which explains why such a large proportion of RNAPII is found in this state. However the longer pausing is related to lower expression [252] since the duration of RNAPII pausing is a limiting step in transcription due to inhibition of productive initiation until either dissociation of the paused complex or release from pausing into productive elongation. The duration of the pause (and thus transcription potential) can be regulated by enhancers [80].

RNAPII also pauses at enhancers, but eRNA transcripts are short (<150bp) and appear to be rapidly degraded by the exosome [93, 238]. The pause site is closer to the enhancer promoter, 35bp downstream, and the frequency of pausing is less than at active genes [238]. Li et al. [135] showed that although 83% RNAPII binding sites are proximal to gene TSSs, 53% of RNAPII chromatin interactions are related to enhancers. Schaukowitch et al. [208] proposed a pause-interfering function for eRNAs, by showing that the eRNAs expressed in neurons were used as decoys for NELF. Knocking down the eRNAs expressed at enhancers compromised pause-release at target genes with consequential decrease in mRNA synthesis.

1.3.2.3 Elongation and Termination

Pause-release into productive elongation is mediated by positive transcription elongation factor b (P-TEFb), which phosphorylates DSIF, NELF, and the C-terminal domain (CTD) of RNAP. This causes NELF to dissociate from the DSIF-NELF complex and triggers the release of RNAPII into productive elongation, as evidenced by the colocalisation of RNAPII, NELF, and DSIF at promoter-proximal regions but RNAPII and

¹⁰Conversion of these cytosines to thymine resulted in higher promoter activity

DSIF only at gene bodies and downstream of their poly(A) sites [192].

Of the RNAPII's that pause, only 7.7% continue to elongation, meaning that roughly 1% of RNAPIIs found at the promoter are likely to complete transcription and produce mature RNA [228].

The speed of elongation is affected by the number and density of exons [111], the cleavage and poly(A) complex (CPA) [176], sequence composition [111]¹¹, as well as the distance from the TSS (rate is slowest at the promoter) [46, 111].

Once elongation begins, the elongating RNAPII complex is generally stable and can transcribe multiple kilobases without dissociating from the DNA [60]. Therefore any changes in post-pausing RNAPII density in gene bodies can be attributed to changes in acceleration of these stable polymerases [111]. Jonkers et al. [111] showed that the acceleration is not constant, and increases as elongation proceeds until a maximum speed is reached¹², before decelerating and pausing near the TES [270]. TES-proximal pausing facilitates sufficient time for cleavage and termination to proceed properly [69, 164, 188], and similarly higher RNAPII signal at exon boundaries suggests that the reduced speed aids spliceosome assembly [176, 126].

Transcription often extends beyond TESs. The rate of elongation determines how far past the termination site the RNAPII will transcribe [39, 164] and can even induce alternative polyA site usage [149, 69]. Schwalb et al. [212] found on average 4 TESs per gene, coinciding with pausing-associated motifs, distributed in a termination window of median width 3300bp. In their study of RNAPII rate mutants, Fong et al. [69] found that slow RNAPII variants terminated further upstream whereas those with faster rates continued downstream suggesting kinetic competition between RNAPII and the Xrn2 exonuclease. Efficient termination is essential for mRNA production since the poly(A) tail influences the stability of mRNA, with longer tails increasing stability [261].

Recalling spatial proximity of TSSs supporting coordinated initiation bursting, the TSS and TES are often found close to one another in 3D space, facilitating RNAP recycling. Nojima et al. [176] showed that all components of the CPA machinery are

¹¹The top 6-mers associated with slowest elongation rates were CpG rich, meaning that gene body CpGs may slow down transcription.

¹²Danko et al. [46] quantified that acceleration increases for the first 15kb.

detected within 500bp of the TSS and suggested a corresponding role in promoter-associated premature termination.

1.3.2.4 Other mechanisms

We define antisense (AS) transcription as transcription from the non-coding strand of the gene of interest, and divergent transcription as promoter-proximal antisense transcription upstream of the TSS. In the original publication describing GRO-Seq, Core et al. [39] highlighted that divergent transcription is prevalent, with most promoters (55% of all genes, 77% of active genes) showing evidence of active RNAPII occupancy within 1kb upstream of the sense TSS peaks. The gene expression, CpG islands, prevalence of pausing, and divergent transcription appear to interact more than would be expected by chance, with divergent transcription marking the most active promoters [53, 39].

Since strand-specific RNAP occupancy maps have become more common, it appears that divergent transcription is prevalent among vertebrates. Promoters of both genes and enhancers often show divergent signal [40, 213, 39], and these bivalent promoters have even been used as markers for enhancer identification [259]. Active enhancers are associated with a bidirectional RNAPII profile and transcription-defined enhancers (those that produce eRNAs) appear to be more likely to be validated as true enhancers (compared to using chromatin modifications alone) [4, 3].

The transcripts resulting from AS transcription are short and unstable (often using cryptic poly(A) sites close to the TSS [176]) and are thus degraded. Although specific genes have well-established evidence of AS regulation [76, 97, 131, 132, 174], the effects of these transcripts is still an active area of investigation with many studies showing little interference with sense transcription [1, 30, 100, 170, 185]. Independent of the transcripts themselves, divergent transcription has been hypothesised to play a permissive role by keeping promoters free of repressive chromatin marks [213].

1.3.3 Techniques to measure engaged RNA polymerase occupancy

Currently there are two main approaches to mapping RNAP enrichment in the genome: chromatin immunoprecipitation (ChIP) based and run-on based. In ChIP-based methods, an antibody is used to select regions of the DNA that associate with RNA polymerases. Since crosslinking between the RNAP and DNA is instantaneous, mapping the RNAP-associated DNA fragments can be used to quantify RNAP occupancy genome-wide through analysis of coverage [162, 260]. Nojima et al. [176] proposed a base-resolution extension, mammalian NET-Seq (mNET-Seq), where DNA and RNA are digested with MNase and the RNAPII-nascent RNA complex is precipitated¹³.

ChIP-based methods have some advantages over methods which quantify nascent transcription¹⁴. ChIP offers the potential of using different antibodies to differentiate between RNAPI, RNAPII, and RNAPIII, and even pull-down different forms of RNAP depending on subunit phosphorylation as above [112, 176]. ChIP can also detect RNAP close to the TSS or in the initiation complex, which cannot be done with the run-on techniques.

However, immunoprecipitation-based methods cannot provide strand information and depend highly on antibody specificity. In contrast, run-on based methods use nascent transcripts as a marker for active transcription. Global run-on sequencing (GRO-seq) was the first iteration of a run-on method mapping active RNAP [39]. Core et al. [39] added a ribonucleotide analogue (bromouridine or BrU) to the nucleotide pool, meaning that transcripts in the process of being synthesised by RNAP would contain a BrU-tag, selected for through immunoprecipitation. By keeping the fragments short (~100bp), they could map the most recently synthesised transcripts in a strand-specific manner.

Precision nuclear run-on sequencing (PRO-Seq) is based on GRO-Seq, but improves on the method by providing base-resolution data through preventing post-incorporation nascent transcript extension using biotin-NTPs instead of BrU. Below is an overview of the protocol (see the full description of the protocol in Mahat et al. [159]):

¹³In their analysis a combination of CTD phosphorylation-specific antibodies).

¹⁴Transcripts that are in the process of being transcribed by RNAP.

1. Nuclei are isolated and the NTP pool is depleted through permeabilization, meaning RNAPs find it hard to continue nascent RNA synthesis due to lack of substrates.
2. During nuclear run-on biotinylated nucleoside triphosphates (biotin-NTPs) are introduced as substrates. In most cases, upon incorporation of bulky biotin-NTPs, RNAPs are prevented from transcribing further¹⁵.
3. The RNA is fragmented since only reads that are long enough to map the 3' end uniquely are required.
4. The first enrichment step filters out fragments lacking a biotin-NTP.
5. 3' adaptors, which are reverse complements of standard 5' adaptors, are ligated before another enrichment step.
6. The 5' end is then de-capped and re-phosphorylated, allowing for 5' adapter ligation at every fragment.
7. The RNA is reverse-transcribed, amplified, and sequenced from the 3' end to ensure that the biotin-NTP is near the 5' (highest quality) end of the reads.

A derivative of PRO-Seq can be used to map active RNAP initiation sites by using the 5' m7G cap as an additional marker. The 5' cap is an inverted nucleotide prepended to nascent transcripts to prevent the 5' end of the transcript from being degraded by exonucleases [112]. In PRO-Cap, the 5' end of fragments are de-phosphorylated first, and then de-capped, to ensure that 5' adapters can only be ligated to the capped fragments. Thus the sequenced reads from PRO-Cap make up a subset of the reads from PRO-Seq. There is some debate regarding when the 5' m7G cap is added. While previous estimates suggest that nascent transcripts as short as 18bp are capped [169], recently Luse et al. [156] calculated this distance to be closer to 30bp. Therefore RNAPs which pause within this 30bp window are unlikely to be pulled down in the PRO-Cap analysis (but may still be included in the PRO-Seq reads).

¹⁵Although considered a bp-resolution method, the chain is not always terminated after incorporation of a biotin-NTP, however the rate of incorporation of more than one is low.

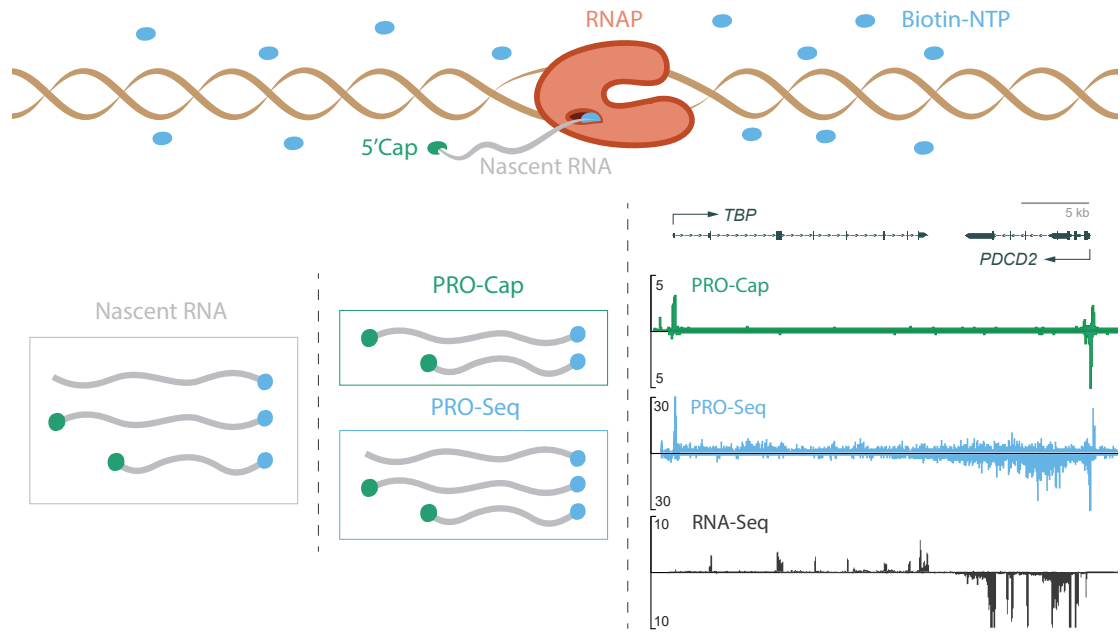


Figure 6: **Schematic of PRO-Cap and PRO-Seq.** RNAPs usually use NTPs as a substrate when synthesising RNA. In these protocols, the NTP pool is depleted and biotin-NTPs are introduced as an alternative for polymerases. Due to the bulky nature of the biotin, the RNAP usually terminates transcription upon incorporation and thus the biotin can be used as a real-time marker of active transcription. In PRO-Cap, the 5' cap is used to map the initiation sites from which the active RNAPs started transcription. Unlike RNA-Seq, PRO-Cap and PRO-Seq reads are not preferentially enriched in exons and often show evidence of antisense transcription.

In our experiments, we used PRO-Seq and PRO-Cap signal to map actively transcribing RNAP position and initiation sites respectively, however many other techniques exist. Zhang et al. [270] proposed a derivative of PRO-Seq which would incorporate a temporal component into the experiments (allowing one to differentiate between fast and slow-moving RNAP). Time variant PRO-Seq (TV-PRO-Seq) uses the saturation of signal across a region to model speed, in other words it measures the signal at different time-points to check how long it takes for a region to be saturated with reads.

Another method, coordinated precision run-on sequencing (CoPRO) [238] can be used to map the 5' and 3' ends of nascent RNAs in a single experiment, combining the data from PRO-Seq and PRO-Cap.

While PRO-Seq leaves the RNAPs to transcribe for 30 minutes, in transient transcriptome sequencing (TT-Seq) [212] this is reduced to only 5 minutes. In contrast to PRO-Seq, TT-Seq measures RNA synthesis rather than RNAs attached to polymerases and due to the short time window, many fragments that would otherwise be degraded (such as enhancer RNAs) can be captured.

Another popular approach is native elongating transcript sequencing (NET-Seq) [?], where the RNAP elongation complex is immunoprecipitated, and the nascent RNA that was being transcribed is extracted and sequenced. In a derivative called mammalian NET-Seq (mNET-Seq), Nojima et al. [176] showed that by combining phosphorylation-specific RNAP antibodies one can differentiate between RNAP modifications across gene bodies¹⁶.

The methods described above do have some general limitations. Cell-to-cell heterogeneity is masked in these bulk analyses, although we can estimate the total signal at a locus over a population we cannot discern whether all cells show similar RNAP occupancy or there exist distinct groups with differing occupancy¹⁷. Similarly the spatial dynamics of transcription cannot be discerned, however imaging-based methods can be used for this purpose [143].

¹⁶They showed that while unphosphorylated CTD RNAP accumulate near the tSS, S2P RNAPs are found in the gene body and near the TES.

¹⁷This is extremely likely, in particular during different stages of replication

1.4 Influence of DNA modifications on transcription

The relationship between DNAmods and transcription is at the heart of understanding how heritable, epigenetic factors can modify gene transcription. Promoter hypermethylation at specific genes has often been synonymous with gene repression [181, 110, 120, 229], but more distal elements also play a role. In studies comparing methylation maps from different tissues, differentially methylated regions are often found in distal regulatory elements, often overlapping enhancers and DNase hypersensitive sites and bound by cell-type-specific transcription factors [22, 95, 225, 13]. Compared to promoters, enhancer hypomethylation can sometimes be more predictive of the expression levels [209]. Hypomethylation should not be synonymous with gene activation as loss of mC can permit transcription repressor binding Wan et al. [247] or induce miRNA regulation of genes [34, 194, 178].

Here we focus on direct (rather than distal) effects of mC and hmC. Many suggest that gene body methylation can repress spurious transcription ([172, 236, 108]). However an additional proposed role is splicing regulation as methylation is often enriched near splice-sites [118, 163, 256, 218, 146, 15].

Focussing on hmC specifically, Ficiz et al. [66] found that a combination of mC and hmC in the promoter is a better indicator of gene expression than 5mC alone, and since promoter hypermethylation is mostly associated with gene repression we can hypothesize that oxidation of mC to hmC could promote transcription by overcoming the repressive properties of mC. hmC enrichment in gene bodies correlates with gene expression as measured by both mRNA levels and RNAP occupancy [54, 144, 165, 120]. For example, in their study of T-cell development, Tsagaratou et al. [241] found hmC enrichment correlated with gene expression throughout the different stages of differentiation. It is important to note that many previous studies describing a link between DNA methylation and expression did not differentiate between mC and hmC¹⁸, so Wen et al. [256]’s finding that a sense strand hmC bias exists (mC (1.0% antisense) and hmC (0.7%sense)) is important as this suggests hmC is more permissive than mC.

¹⁸Both are not converted in bisulfite sequencing, thus most 450k array or BS experiments would group them.

An alternative role was proposed by Hahn et al. [84] who suggested that gene body hmC on the sense strand may prevent antisense transcription.

We know that elongation rate has a strong effect on transcript length, and stability. DNA modifications have been linked to RNAP pausing Danko et al. [46], Veloso et al. [244], but there is some debate in the field. Lorincz et al. [152] suggest that methylation encourages a closed chromatin conformation in intragenic regions, and thus “reduces the efficiency of RNAPII elongation”. However, other studies have found no relationship between mC and RNAP rate [249, 106]. Although the rate of elongation has also been inversely related to abundance of fC, caC, and N6-mA, the same was not the case for hmC but there are very few studies analysing this relationship [249, 250].

1.4.1 Aim of thesis

In this project we aim to elucidate how DNA modifications affect gene expression through their effect on actively transcribing RNAP. The relationship between RNAP occupancy and gene expression is clear as high expression requires consistent mRNA synthesis¹⁹.

As mentioned in the previous section, many studies have shown a relationship between DNA modifications and expression. The repressive functions associated with mC in particular can explain the association between promoter hypermethylation and gene expression, but the role of gene body hypermethylation in highly expressed genes has thusfar been quite difficult to explain. Therefore we combine multiple data types, representing measures of DNA modifications (TAPS and TAPS- β), active transcription (PRO-Cap and PRO-Seq), and gene expression (RNA-Seq) to isolate the effects of mC and hmC on active RNAP and gene expression (Figure 7).

We decided to take a genome-wide approach to the question, and thus evaluate the effect of global mC and hmC changes on transcription and expression. In Chapter 3 we induce genome-wide perturbations to mC and hmC. We then map RNAP in the modified cells as a measure of active transcription in Chapter 4. We combine expression

¹⁹This relationship is not homomorphic and thus the opposite is not always true, namely high occupancy does not imply high expression. For example, active enhancers are often transcribed and show high RNAP occupancy but do not produce mature mRNAs.

with RNAP occupancy and DNA modifications in Chapter 5 to see the strength of any mediating effects.

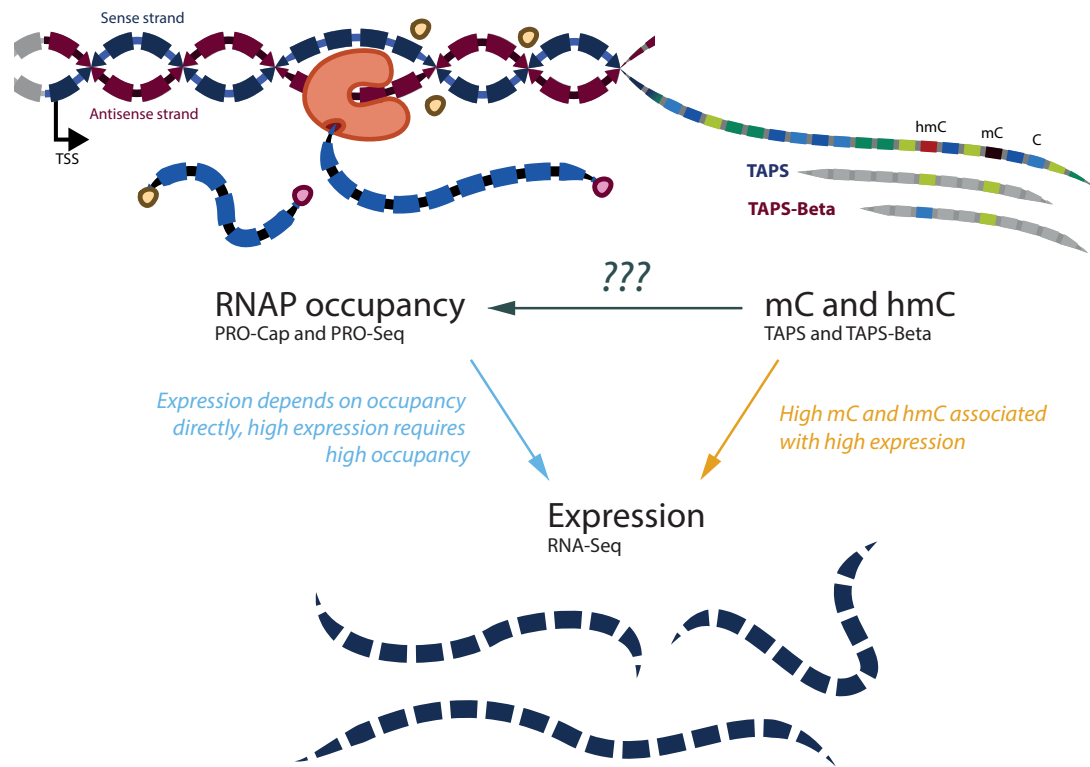


Figure 7: **Aim of thesis.** Isolate how hmC and mC affect RNAP, and how the changes in transcription affect gene expression.

2 Methods

2.1 Annotation

2.1.1 Genome annotation

The gencode.v19.annotation.gtf.gz file was downloaded from <https://www.encodeproject.org/files/gencode.v19.annotation/> and subsequently loaded into R. We kept HA-VANA annotations only, and created a gene annotation (`type == "gene"`) and a transcript annotation (`type == "transcript"`) to be used in the analyses.

2.1.1.1 Metagenes

To make the metagene plots, the 4kb region upstream, the gene body, and the 4kb region downstream of the annotated genes were split into 100 windows using `bedtools makewindows -n 100`, resulting in 300 regions per gene.

For the TSS and TES metaplots, we selected protein-coding genes which did not overlap with any other protein-coding genes within 4kb (i.e. the gene boundaries are at least 8kb apart). This ensured that the meta-TSS and meta-TES plots would not contain signal from adjacent genes. Any genes shorter than 2.5kb were then removed, this left 6,628 genes to draw the boundary metagene plots from. We increased granularity for these figures and focused on the 2kb regions flanking the TSS and TES, split into 50bp bins.

2.1.2 CpGI annotation

CpG island annotation (hg19) was extracted directly from using the `annotatr` package [32]. In order to draw the summarised plots, for each CpGI we defined the up- and downstream 2kb region as the shore, and the regions a further 2kb up- and downstream as shelves (all CpGIs were included, but any overlapping shores or shelves were excluded). Each region was then split into 20 bins, in which modification levels were calculated:

```
bedtools map -a CPGI_BINS -b MOD_BG -prec 3 -c 4 -null 'NA' -o mean, count
```

2.1.3 Enhancer annotation

Enhancers and promoters can be distinguished by the methylation on H3K4, with H3K4me1 enriched in enhancers and H3K4me3 marking promoters. An enhancer can then be classified as active if H3K27ac co-occurs with the mono-methylation, primed if the only mark is H3K4me1, or poised if also marked by a repressive H3K27me3 modification [43, 28]. There is some debate regarding these classifications [93, 51, 198], and since we didn't generate data directly we used multiple annotation resources.

2.1.3.1 FANTOM5

The `annotatr` package [32] provided enhancer coordinates from the FANTOM5 project (phase 2.0) [175]. We downloaded the latest coordinate and count data directly from FANTOM5²⁰. The consortium generated bp-resolution profiles of transcription start sites (profiled using CAGE) across 432 primary cell, 135 tissue and 241 cell line samples which were then used by Andersson et al. [3] to analyse enhancer transcription (eRNA) as a proxy for active enhancer activity. The data consists of 65,423 enhancer coordinates, which we supplemented with the HEK293 transcripts per million (TPM) data to define 6,952 active enhancers with non-zero TPM.

2.1.3.2 EnhancerAtlas v2.0

Gao and Qian [73] created the EnhancerAtlas v2.0 database to model enhancer-gene interactions. The predictions are based on multiple data types, including H3K4me1/H3K27ac ChIP, DNase-seq/ATAC-seq accessibility data, RNAPII occupancy, CAGE and GRO-Seq transcription, and chromatin interaction ChIA-PET. The enhancer map²¹ for HEK293 cells show 119,167 enhancers associated with 12,282 genes, however there was no associated data source map available for this cell line.

²⁰Retrieved from http://fantom.gsc.riken.jp/5/datafiles/phase2.6/extra/Enhancers/human_permissive_enhancers_phase_1_and_2.bed.gz and http://fantom.gsc.riken.jp/5/datafiles/phase2.6/extra/Enhancers/human_permissive_enhancers_phase_1_and_2_expression_tpm_matrix.txt.gz

²¹Downloaded from http://www.enhanceratlas.org/data/AllEPs/hs/HEK293_EP.txt

2.1.3.3 Enhancer classification

Based on publicly available histone modification ChIP data, we defined our own set of HEK293 enhancers. This allowed us to check for consistency between the different annotations and also discern between different types of enhancers. To build the custom annotation for HEK293 cells, H3K4me3 (ENCFF610MBI), H3K27ac (ENCFF699JDN), and H3K4me1 (ENCFF734LIC) ChIP peaks were downloaded from the ENCODE database (www.encodeproject.org) and a H3K27me3 map from www.chip-atlas.org (SRX6369406). The bigBed files were loaded into R using `rtracklayer::import.bb()`. The H3K4me3 data was mapped to hg38, so we used `rtracklayer::liftOver()` to ensure compatibility with hg19. The resulting hg19 peaks were only split in 0.1% of cases and the width distribution remained equivalent to the hg38 map, meaning that most of the regions could easily be converted between coordinate systems.

2.1.3.4 Enhancer annotation comparison

We calculated overlaps between the different annotations using `GenomicRanges::findOverlaps()` and saw significant variability between the annotation databases (Figure 8). Only 13.3% of the active FANTOM enhancers and 9.5% of the EnhancerAtlas enhancers overlapped with H3K4me1 peaks, respectively. There was little direct overlap between the datasets as well (0.6%-0.7% depending on the reference). The peaks in the EnhancerAtlas dataset were enriched for CpG islands and promoters in particular, suggesting a dominance of promoter-promoter interactions in the data.

2.2 DNA Modifications

2.2.1 Pre-processing of TAPS and TAPS- β data

The quality of the raw reads was first analysed using FastQC v0.11.2 [187], and then trimmed using trimgalore [123] with `-q 20 --length 20`, to remove any low quality and/or short reads. The trimmed .fastq files were aligned using `astair align` [150], against a combined reference containing both hg19 sequences and those associated with each spike-in. Any duplicates were removed using `picard MarkDuplicates` [101] with `REMOVE_DUPLICATES=true`.



Figure 8: **Overlap between enhancer annotation.** Heatmap showing the percentage of reference peaks which overlap the candidate peaks.

To analyse conversion, we extracted 4 subsets from each .bam file, containing reads that mapped to each of the 4 spike-ins. We then ran `astair call` at each C (`--context all`), considering reads with quality 13 or more (`--minimum_base_quality 13`), based on the TAPs conversion method (`--method mCtoT`). Since the 250mer and 160mer were very highly covered, we increased the maximum allowed depth (`--max_depth 4000000`).

To parallelise, we called modifications on each chromosome separately, but in a CpG context only (`--context CpG`) since we saw little conversion outside a CpG context, these were then combined to form one file per sample. The files were filtered to include only non-SNP CpGs with at least 5 reads. We generated 2 bedgraph files from each filtered mods file: one with the β value, i.e. $\text{MOD}/(\text{MOD} + \text{UNMOD})$, as the score, and the other with the coverage.

To evaluate hmC levels and changes in DNA modifications between conditions, the β bedgraphs were then intersected with `bedtools intersect` [190] so that changes

Ref	Other	Interpretation
TAPS (uns)	TAPS- β (uns)	hmC (uns) - base hmC level in unstimulated HEK293 cells
TAPS (Aza)	TAPS- β (Aza)	hmC (Aza) - hmC level after Aza treatment
TAPS (Dox)	TAPS- β (Dox)	hmC (Dox) - hmC level after Dox treatment
TAPS- β (uns)	TAPS- β (Dox)	mC loss (Dox)
TAPS- β (uns)	TAPS- β (Aza)	mC loss (Aza)
TAPS (uns)	TAPS (Dox)	uC gain (Dox) - Complete demethylation after Dox
TAPS (uns)	TAPS (Aza)	uC gain (Aza) - Complete demethylation after Aza
hmC (uns)	hmC (Dox)	hmC gain (Dox)
hmC (uns)	hmC (Aza)	hmC gain (Aza)

Table 1: Interpretations of $\beta_{ref} - \beta_{other}$ intersections.

in β level could be calculated per CpG (as long as the CpG is covered by at least 5 reads in each sample). We first calculated the hmC level per CpG in each condition by subtracting the mC level (TAPS- β) from the mC+hmC level (TAPS). The remaining measures, i.e. mC loss, unmodified C (uC) gain, and hmC gain, were calculated as shown in Table 1.

2.2.2 Modification level simulation

To investigate the cause of negative modification changes, i.e. cases where the TAPS- β modification level exceeds TAPS for example, we used the coverage distributions and mean conversion rates per CpG to simulate the expected β frequencies given a simplified model.

Given a CpG with M modified and U unmodified reads, the β value depends on the coverage, $N = M + U$, and conversion efficiency. If the conversion efficiency is below 100% and there is no bias in terms of context, a fully methylated position is unlikely to be called as fully methylated ($\beta = 1$) when the coverage is high since the probability of incomplete conversion increases with each read proportionally to the false negative rate. Assuming independence between reads and a conversion efficiency of 95%, we can model the expected number of modified positions, $\hat{M}$, using a binomial distribution with M trials and success probability 0.95.

To model the hmC level in the unstimulated sample, we simulated 100,000 CpGs, $c_i, i = 1, \dots, 100000$, based on the coverage distributions in our data and some simplifying assumptions. For each CpG:

- The true mod (mC+hmC) level, mod_{true_i} , was binary (0 or 1), with a probability of a position being modified, $p_{mod} = 0.7$.
- Since the hmC level in the unstimulated sample was extremely low, we decided to model the proportion of hmC, hmC_{true_i} , using an exponential distribution with mean, $p_{hmC} = 0.0001$. This meant that the level of mC was given by $mC_{true_i} = mod_{true_i} - hmC_{true_i}$.
- We simulated the corresponding coverage values for the simulated modification and mC levels based on coverage frequency in the unstimulated TAPS (N_{TAPS_i}) and TAPS- β ($N_{TAPS-\beta_i}$) samples, respectively.
- Working backwards, the MOD counts in TAPS and TAPS- β were calculated as $MOD_{TAPS_i} = N_{TAPS_i} \times mod_{true_i}$ and $MOD_{TAPS-\beta_i} = N_{TAPS-\beta_i} \times (mod_{true_i} - hmC_{true_i})$ (in both cases rounded to the nearest whole number).
- Using the simulated counts, we then "converted" each mC using a binomial distribution with 0.95 success probability, and each hmC with a 0.75 success probability (based on the average conversion rates from the spike-ins). Thus, the simulated conversion rate in the TAPS sample at c_i is given by $\text{rbinom}(1, MOD_{TAPS_i}, 0.95) + \text{rbinom}(1, MOD_{TAPS_i} - MOD_{TAPS-\beta_i}, 0.75)$.

The effect of random, genome-wide changes was simulated similarly, with the exception that the mC level of the +5-aza-CdR was directly related to that of the unstimulated sample. We modelled the demethylating effect of the treatment to have a $\max(0, N(0.25, 0.1))$ distribution, such that on average the treatment would remove 25% of the methylation per CpG (but would never add methylation).

To compare unstimulated and +5-aza-CdR TAPS- β samples:

- The true mod level in the unstimulated sample, mod_{uns_i} , was binary (0 or 1), with a probability of a position being modified, $p_{mod} = 0.7$.
- The true mod level in the +5-aza-CdR was calculated as $mod_{aza_i} \times (1 - x_i)$, $x_i \in \max(0, N(0.25, 0.1))$.

- As before, we simulated a coverage value based on coverage frequency in the unstimulated TAPS- β sample, N_{unsi} , and another based on the +5-aza-CdR TAPS- β sample, N_{aza_i} .

Incorporating the conversion efficiency at mC as in the hmC analysis pipeline, we calculated a simulated modification level in both the unstimulated and +5-aza-CdR samples.

2.2.3 Metagenes

The bedgraphs containing β values, and changes (the gains and losses) were summarised across the metagene annotations described in section 2.1.1.1 using `bedtools map (-c 4 -null 'NA' -o mean)`.

2.2.4 Genome Browser

The β values were extracted as the score column for .bedgraph files, and corresponding .bw files were generated using `bedGraphToBigWig` [117] and visualised in the Integrated Genome Browser [70].

2.2.5 Differential methylation analysis

For each CpG, we used the modification level and the coverage to ran a proportion test (`prop.test()` from the `stats` package [191]) to define significance of changes. This assumes that CpGs with highest coverage are more informative at showing true changes in DNA modifications. Since we expect to have positive values when calculating hmC level per CpG, the loss in mC or gain in C after demethylating treatments, we used a one-sided test because negative values do not make sense in these contexts.

2.3 mRNA

2.3.1 Pre-processing of mRNA data

There were 4 lanes per sample, and, prior to merging the files, we analysed the individual FastQC [5] profiles of the lanes separately and found a clear outlier in the +5-aza-CdR

experiments. The R2 reads from lane 1 were of bad quality, caused by problems with the flow cell based on the *Quality Per Tile* output from FastQC (Figure 9). This lane was left out, and the remaining 3 were merged for each sample using `cat`.

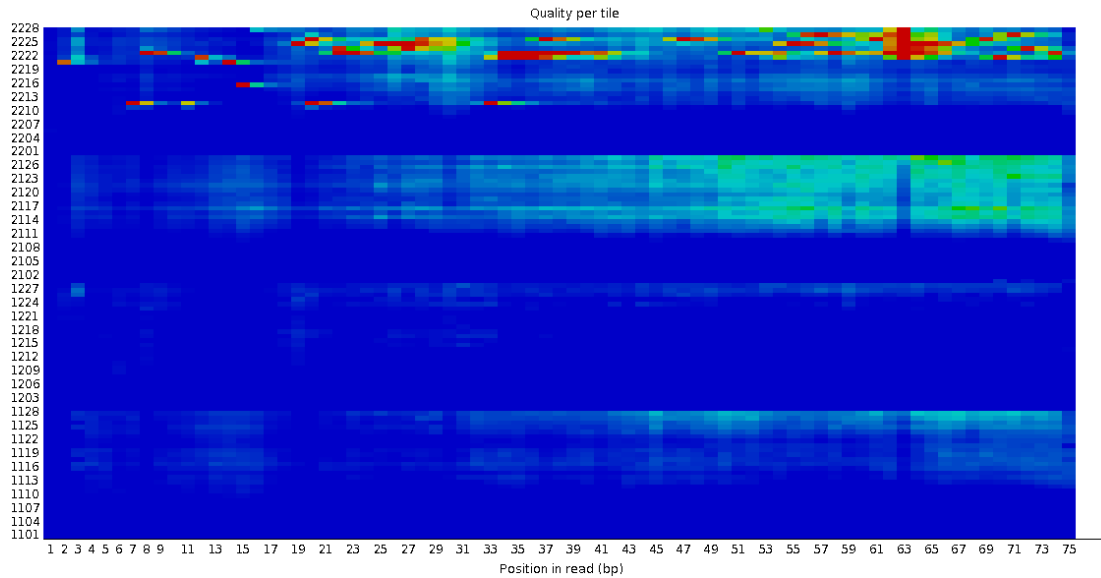


Figure 9: **Representative tile quality in Lane 1, read 2.** The Quality Per Tile module output from FastQC for sample WTCHG_396576_210178_2.fastq.gz (one of the Aza_1 lanes). The clear banding shows that sequence quality reduced significantly for a fraction of tiles as the read length increased.

To quantify transcription, we used `salmon quant` [184]. The software uses quasi-mapping to estimate the number of reads mapping to the transcriptome while accounting for various biases known to be common in RNA-Seq datasets including GC-bias and sequence bias [154], as well as normalising for length based on the fragment size distribution [201]. To evaluate the effects of different normalisation strategies, we used combinations of the `--gcBias` and `--seqBias` parameter options (see Section 5.2.1).

2.3.2 Differential expression analysis of mRNA count data

The outputs from `salmon quant` were imported using `tximport::tximport` [221]. The differential expression analysis was carried out using DESeq2 [153] on transcripts with 10+ reads only, and batch number in the design formula to reduce any batch effects:

```
dds.txi.g <- DESeqDataSetFromTximport(txi.g, colData = colDataGroup,
                                     design = ~ Sample + Condition)

keep <- rowSums(counts(dds.txi.g)) >= 10
dds.txi.g <- dds.txi.g[keep,]
dds.txi.g <- DESeq(dds.txi.g)
dds.res <- DESeq2::results(dds.txi.g, contrast = contrastVec, alpha = 0.05)
```

2.4 Engaged polymerase occupancy

Please note: To avoid including too much code here, additional code has been included in Appendix 1.

2.4.1 QC analysis and deduplication

The PRO-Cap fastq.gz files were first trimmed to remove adapters²² with `trimgalore -q 20 --length 20` [123]. We then used `umitools` [220] to cleave off a 5' 4bp UMI and a 3' 4bp UMI and append this 8bp combined UMI to the read names

```
umi_tools extract \
    --bc-pattern="^(?P<umi_1>.{4}).*(?P<umi_2>.{4})$" \
    --extract-method=regex
```

The resulting reads were aligned with STAR [50] to a custom hg19-galgal5 genome (see Section 2.4.2), STAR `--outFilterMultimapNmax 1`. Although we expected most of the reads to be unspliced, using a splicing-aware aligner seemed more sensible in case of mRNA contamination. We split the PRO-Cap bam files into *long*, those that were 63bp or more (presumed adapter content 5bp or less since the 8bp of the UMIs have been removed), and *short*, those length 12-62 using `samtools` [136]:

```
samtools view -h trimmed_cleaved_aligned_sorted.bam | \
    awk -v var=63 'length($10) < var || $1 ~ ^@/' | \
    samtools view -bS - > trimmed_cleaved_aligned_sorted_maxLength_62.bam
```

²²The trimming also removes low quality reads, but based on the FastQC results, there were very few low quality reads in this data.

For the *long* files, the combination UMI is trimmed back to 4bp since the 3' 4bp sequence is more likely to be genomic rather than the 3' 4bp UMI:

```
samtools view -H trimmed_cleaved_aligned_sorted_minLength_63.bam > header.sam

samtools view trimmed_cleaved_aligned_sorted_minLength_63.bam | \
    awk '{ $1 = substr($1, 1, length($1) - 4); print }' OFS='\t' | \
    cat header.sam - | \
    samtools view -b - > trimmed_cleaved_aligned_sorted_minLength_63_cropped.bam
```

The resulting *short* files (with 8bp combination UMI) and *long* files (with 5' 4bp UMI) are then deduplicated with `umi_tools dedup` and subsequently merged together with `samtools merge`.

The PRO-Seq fastq.gz files were trimmed to remove the 3' 4bp UMI before mapping:

```
trimgalore -q 20 --length 20 --three_prime_clip_R1 4
```

The 5' 8bp UMIs were then extracted and appended to the read names with `umitools`, `umi_tools extract --bc-pattern=NNNNNNNN`. Alignment and deduplication was the same as for PRO-Cap. In their GRO-Seq analysis, Bi et al. [18] allowed a maximum of 3 reads at each genomic position to ensure the data wouldn't be biased by high pausing. However in our analyses, looking for changes between conditions, this did not seem sensible due to high similarity between treatments and the previously mentioned non bp-resolution nature of the experiments.

2.4.2 The reference genome

Chicken cells were added to the HEK293 cells to aid normalisation in the event of genome-wide increases or decreases resulting from the treatments. We downloaded the galGal5 sequence from UCSC, <ftp://hgdownload.soe.ucsc.edu/goldenPath/galGal6/bigZips/galGal6.fa.gz> and created a custom hg19_galgal6 genome for this analysis:

```
bioawk -c fastx '{ print ">m." $name "\n" $seq }' galGal6.fa.gz > galGal6_SPIKEIN.fa
cat $FASTA_HG19 galGal6_SPIKEIN.fa > hg19_galgal6.fasta
```

The reads were aligned against this genome, and the deduplicated bam files could then be split into reads mapping to hg19 and galgal5, respectively.

2.4.3 Processing the .bam files

2.4.3.1 Generation of signal tracks

We extracted the strand-specific start positions from the PRO-Cap and PRO-Seq hg19 reads, with one read representing a single molecule with an active RNAP²³. The data was then merged into bedgraph files, with the score representing the number of unique RNAPs at a position. To combine tracks we used `bedtools unionbedg` and samples were normalised using counts per million (CPM). We used `bedGraphToBigWig` [117] to generate bigwig tracks as before. For more details, see Appendix 1 for implementation.

2.4.3.2 Downsampling PRO-Cap and PRO-Seq data

Based on the spike-ins, the treatments did not induce genome-wide changes in engaged RNAP occupancy, but the samples had slightly different depths. Therefore, to facilitate direct comparison of samples in some of the analyses, we downsampled the PRO-Cap .bam files to 23 million reads and PRO-Seq to 53 million reads (guided by the counts in the smallest samples), see Appendix 1 for implementation.

2.4.3.3 Metaplot generation

See Appendix 1 for implementation.

2.4.4 Identifying regions of polymerase initiation

2.4.4.1 The HMM

²³Recall that in PRO-Cap the sequencing starts at the 5' cap, but in PRO-Seq the sequencing starts at the 3' end and on the opposite strand. Therefore we are interested in the 5' base in both but the strand must be changed in the PRO-Seq.

2.4.4.1.1 Counts, observed variables, and blocking

Based on the PRO-Cap profiles usually showing initiation across multiple adjacent bases, we decided to combine the data across 10bp bins in an effort to reduce computational burden. The signal and peaked-ness value associated with each 10bp bin are then calculated as these are the columns used as input to the HMM. Based on this data we also define stretches of DNA with non-zero PRO-Cap density on which to run the peak-caller (to avoid the computational burden of running a HMM on the entire chromosome). See Appendix 1 for implementation.

2.4.4.1.2 Transition matrices, peak calling, and trimming

Although we pre-filtered long stretches that are empty through blocks, the probability of starting in a background state is still highest:

$$\text{Init.} = \begin{matrix} & & \text{Init.} \\ & B & \left(\begin{array}{c} 0.99 \\ 0.001666667 \\ 0.001666667 \\ 0.001666667 \\ 0.001666667 \\ 0.001666667 \\ 0.001666667 \end{array} \right) \\ & BP1 & \\ & BP2 & \\ \text{Init.} = & BP3 & \\ & NP1 & \\ & NP2 & \\ & NP3 & \end{matrix}$$

Following Core et al. [40], we expected most background regions to be empty, and the enriched regions in narrow peaks to be surrounded by depleted signal whereas the broad peaks assumed high enriched proportions:

$$\begin{array}{c}
\begin{array}{c}
\text{empty} \quad \text{depleted} \quad \text{enriched} \\
B \\
BP1 \\
BP2 \\
BP3 \\
NP1 \\
NP2 \\
NP3
\end{array}
\end{array}
\begin{pmatrix}
0.9 & 0.05 & 0.05 \\
0.09 & 0.01 & 0.9 \\
0.09 & 0.01 & 0.9 \\
0.09 & 0.01 & 0.9 \\
0.1 & 0.45 & 0.45 \\
0.1 & 0.45 & 0.45 \\
0.1 & 0.45 & 0.45
\end{pmatrix}$$

The apex of any narrow peaks was expected to fall into the NP2 class:

$$\begin{array}{c}
\begin{array}{c}
\text{nopeak} \quad \text{peak} \\
B \\
BP1 \\
BP2 \\
BP3 \\
NP1 \\
NP2 \\
NP3
\end{array}
\end{array}
\begin{pmatrix}
1 & 0 \\
1 & 0 \\
1 & 0 \\
1 & 0 \\
1 & 0 \\
0 & 1 \\
1 & 0
\end{pmatrix}$$

In contrast to Core et al. [40], we decided to reduce the self-transition probability between background states from 0.99 to 0.9 in order to pick up smaller peaks, the remainder of the matrix remained unchanged:

$$T = \begin{matrix} & \begin{matrix} B & BP1 & BP2 & BP3 & NP1 & NP2 & NP3 \end{matrix} \\ \begin{matrix} B \\ BP1 \\ BP2 \\ BP3 \\ NP1 \\ NP2 \\ NP3 \end{matrix} & \begin{pmatrix} 0.9 & 0.05 & 0 & 0 & 0.05 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0.5 & 0.5 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0.5 & 0.5 & 0 \\ 0 & 0 & 0 & 0 & 0.45 & 0.1 & 0.45 \\ 0.5 & 0 & 0 & 0 & 0 & 0 & 0.5 \end{pmatrix} \end{matrix}$$

We implemented the HMM model with the above parameters in python using the pomegranate package [210] and then trimmed as described in Section 4.2.8.2. See Appendix 1 for implementation.

2.4.5 Linking initiation regions to features

In order to assign genes to initiation peaks as described in Section 4.2.10.2, we extensively used the `plyranges` [130] package. See Appendix 1 for implementation.

2.4.6 Identifying regions of polymerase elongation

To define elongation regions as described in Section 4.2.10.3, we first identify major initiation regions for each gene.

2.4.7 Selecting major initiation regions for each gene

See Appendix 1 for implementation.

2.4.8 Assigning probable elongation regions

See Appendix 1 for implementation.

2.4.9 Transcription statistics

2.4.10 Uniformity index to evaluate stalling

In their analysis of intron retention, Li et al. [137] measure the “flatness” of the read distribution in a genomic region using entropy as follows:

- The region r_i is divided into 8 bins, the reads in each bin are counted, $c_i = [c_{i,1}, \dots, c_{i,8}]$.
- The probability mass function is then: $p_{i,j} = \frac{c_{i,j}}{\sum_{k=1}^8 c_{i,k}}$ and the entropy $E_i = -\sum_{j=1}^8 p_{i,j} \log_2(p_{i,j})$
- The maximum entropy of 8 bins is 3, since the least informative distribution for prediction would be uniform, i.e. if each of the 8 bins was equally full then the $p_{i,j} = 0.125$ for all bins j ²⁴. So the entropy is then normalised to be between 0 and 1 by calculating $E_i/3$.

In their analysis, Li et al. [137] split the introns into 8 bins only, but since elongation regions are generally longer than introns we decided to bin them into 20 regions instead (with a maximum entropy of 4.321928). See Appendix 1 for implementation.

2.4.11 Termination index

See Appendix 1 for implementation.

2.4.12 Differential RNAP initiation and occupancy analysis

We used DESeq2 [153] to calculate which initiation peaks and elongation regions showed differences upon +5-aza-CdR and +Dox treatment. Using un-normalised counts for each of the 9 replicates separately, we compared the effect of +5-aza-CdR and +Dox treatment against the unstimulated cells. We accounted for batch in our design matrix and since some of the changes were difficult to see in the genome browser, we included an additional threshold ($|LFC| > 1$) in some analyses for robustness. We tried multiple combinations of DESeq2 parameters, including testing the effects of the independent-Filtering parameter and apeglm shrinkage but found that these increased replicate vari-

²⁴Thus $E_i = -1 \times 8 \times (0.125 \times \log_2(0.125)) = 3$

ability and thus we decided to use no independentFiltering and Benjamini-Hochberg p-value correction. See Appendix 1 for implementation.

3 Induction of unbiased genome-wide DNA modification changes

3.1 Introduction

DNA modifications have a role in transcription regulation, but this role is complex. Although $\sim 70\%$ of CpGs in the mammalian genome are methylated, promoter regions and CpGI's are generally left unmethylated, allowing for transcription factor binding. Surprisingly, while promoter-proximal methylation is associated with gene repression [181, 110, 120, 229], multiple studies have found a *positive* correlation between gene body mC and hmC and expression [54, 144, 165, 120, 241], suggesting that DNA modifications may interact with RNA polymerase (RNAP) during transcription. hmC is roughly 10-fold less abundant than mC, and its specific impact is poorly understood, though some studies have highlighted links with transcription. Janoušková et al. [106] used fully modified templates and the Pveg promoter and found that hmC-modified DNA transcription was higher by a factor of 3.5 compared to unmodified or methylated DNA which suggested a transcription boosting role for hmC. With regards to regulation, Wen et al. [256] found enrichment of hmC at 5' splice-sites, poised (compared to active) enhancers, as well as a hmC sense-strand bias compared to mC in gene bodies.

In this project, we want to elucidate and isolate the effects of mC and hmC on active transcription using a system that affects mC and hmC levels genome-wide. By modifying DNA modifications genome wide we could check whether the effects are common, and if there are additional features governing gene regulation by DNA modifications. This would give more power to detect changes in genes of different length, CpG density, and expression level. Whereas genome-wide *hypermethylation* is comparatively rare²⁵, genome-wide *hypomethylation* is a hallmark of cancers [56], therefore a demethylating system was preferable in terms of both ease²⁶ and biological interpretation.

In this chapter we describe a suitable system of DNA modification perturbation and check whether it works effectively and in an unbiased manner, providing a platform to

²⁵Although can occur, with examples in IDH-mutant acute myeloid leukemia for example [258].

²⁶Induction of mC in an unbiased manner genome-wide is more difficult.

evaluate the role of DNA modifications on active transcription and mRNA abundance (Chapters 4 and 5). We begin by describing the experimental set-up, designed by Dr. Pijus Brazauskas [27]. The quality metrics required for mC and hmC analyses, including conversion efficiency and coverage, are then discussed in detail. We describe how the modification levels and induced changes are estimated, and expand on problems with such estimations. Finally, we analyse mC and hmC levels in functional regions, i.e. genes and CpGIs, to check magnitudes of changes in the genomic regions we are most interested in and for comparison with published results (considering TAPS and TAPS- β are relatively new techniques).

3.2 Results

3.2.1 Experimental Conditions

The currently accepted view is that the vast majority of hmC originates from enzymatic oxidation of mC (Figure 10). Consequentially, given a constant production of mC we expect that increase in hmC will result equivalent loss of mC, thus any changes resulting from excess hmC may be a consequence of having 1) more hmC or 2) less mC. To deconvolute the effects of demethylation (loss of mC only) and C gain (loss of mC and hmC), we modified conversion rates within the active demethylation pathway (Figure 10) to induce genome-wide changes in mC and hmC such that the effects of hmC gain could be isolated from those of mC loss.

3.2.1.1 TET1-CD upregulation via Dox induction

TET enzymes are responsible for mC to hmC conversion, thus by upregulating their expression we can increase the abundance of hmC. Our aim was to induce hmC in an unbiased manner genome-wide (rather than the function of TET), therefore using the full-length protein would not achieve the desired uniform distribution of changes due to targeted oxidation. When comparing the effects of the full-length TET1 to the TET1 catalytic domain (TET1-CD) in HEK293T cells, Jin et al. [107] found that upregulation of TET1-CD induced global DNA demethylation genome-wide with strong differences at CpG islands between the full length and TET1-CD profiles. Namely, rather than

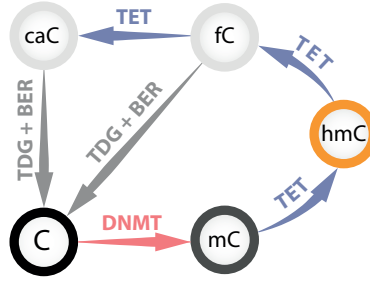


Figure 10: **Active demethylation pathway.** Unmodified cytosine (C) is methylated by DNA methyltransferases (DNMT's). The TET enzymes oxidise mC to hmC, fC, and caC. The latter, fC and caC, are unstable and removed by thymine DNA glucosylase as part of the base excision repair (BER) pathway.

decreasing methylation levels throughout the genome, the full length enzyme remains near CpG islands, preventing methylation from spreading into these regions. Therefore, HEK293 and U2OS cells were modified to contain a Dox-inducible TET1 catalytic domain (TET1-CD) [27].

3.2.1.2 Demethylation with 5-aza-CdR

To achieve equivalent genome-wide mC loss we reduce DNMT activity. 5-aza-2'-deoxycytidine²⁷ (5-aza-CdR) is a nucleotide analog which, when incorporated into DNA by DNA polymerases, depletes the DNMT pool through covalent binding [36]. When DNMT1 attempts to copy methylation marks onto a newly synthesised daughter strand after replication, if the base to methylate is a 5-aza-CdR then the protein cannot dissociate from the DNA and is prevented from methylating other positions. Robert et al. [200] showed that depletion of DNMT1 (but not 3A and 3B) affects the ability of 5-aza-CdR to reactivate genes, thus this treatment affects mostly the maintenance enzyme and passively reduces the level of mC genome-wide since the number of active DNMT1s drops with each replication cycle.

²⁷Also known as the drug decitabine.

3.2.1.3 Isolating the effects of 5hmC induction from those of demethylation

The next step is to ensure that the level of mC loss after TET1-CD upregulation through Dox induction (+Dox) is equivalent to the mC loss after +5-aza-CdR treatment. First, HEK293 cells were treated with 1 μ g/ml Dox for 48 hours. Compared to unstimulated HEK293 cells, the mC (calculated as a percentage of all N) dropped by over 20% from 0.77% to 0.6%, with hmC gain accounting for 47.1% of the difference (the rest being completed active demethylation) (Figure 11 and Table 2).

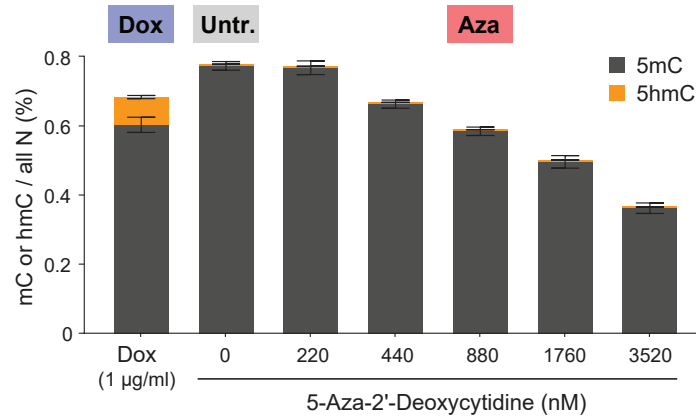


Figure 11: **Levels of mC and hmC in HEK293 cells measured by mass spectrometry.** Bar plots showing the titration curve of the percentage of mC and hmC (measured by mass spectrometry) as a proportion of all nucleotides in HEK293 cells after 48 hours of treatment with 1 μ g/ml Dox or multiple concentrations of 5-aza-CdR. All measurements done by Dr. Paolo Spingardi.

	<i>mC / all N (%)</i>		<i>hmC / all N (%)</i>	
	average	SEM	average	SEM
unstimulated	0.77	0.01	0.005	0.0001
+Dox (1 μ g/ml)	0.60	0.02	0.08	0.005
+5-aza-CdR (880nM)	0.58	0.01	0.005	0.0004

Table 2: **Levels of mC and hmC in HEK293 cells measured by mass spectrometry.** All measurements done by Dr. Paolo Spingardi.

To match the mC loss between +Dox and +5-aza-CdR, a titration experiment was carried out. The final level of mC after +5-aza-CdR treatment is dependent on both the concentration and time of exposure and, keeping the time of exposure consistent at 48

hours, the cells were treated with 5-aza-CdR at concentrations of 0nM, 220nM, 440nM, 880nM, 1760nM, and 3520nM. Based on the titration curve, cells should be treated for 48 hours with 880nM of 5-aza-CdR to achieve mC loss equivalent to 48 hours of 1 μ g/ml Dox induction (Figure 11).

3.2.2 Detection of base-resolution DNA modification changes in HEK293 cells

3.2.2.1 Quality control

3.2.2.1.1 Conversion rate of spike-ins

Our goal is to accurately measure mC and hmC levels at base resolution, however the complexity of the TAPS and TAPS- β protocols means that there is room for experimental variability to bias the results. We can detect problems and inconsistencies in conversion by adding DNA with known modification levels (a spike-in) to the samples as a control, to check how accurately the method detects modifications in each experiment and thus ensure that the data is comparable and reproducible. The conversion rates calculated based on spike-ins can provide a measure of false negatives, i.e. how often a modification is missed, and false positives, i.e. how often we incorrectly identify a base as modified. Without them we cannot know how well the experiment worked, or whether the samples are comparable²⁸.

For each experiment, we need measures of conversion at C, mC, and hmC in various contexts. To this end, the 3 spike-ins used by [150] are added to each sample with known modification rates, namely:

- To estimate mC conversion rate: unmethylated lambda-phage DNA (Promega) which has been methylated at each CpG using MSssI (NEB).
- To estimate false positive rate: unmethylated 2_kb amplicon from the pNIC28-Bsa4 plasmid (Addgene, catalog no. 26103).

²⁸If samples A and B had the same base mC levels but B had 20% worse conversion, then without accounting for the difference in conversion we would incorrectly assume that B had 20% less mC than A

- To analyse context-specific conversion: a synthetic 160mer with **NmCNN** and **NhmCNN**.

The information from these spike-ins was insufficient to fully understand conversion efficiency as only the 160mer contained hmC, and not in the most biologically relevant context. Therefore we designed a synthetic spike-in with the following requirements:

- The spike-in contains many positions which are hemi-hydroxymethylated/hemi-methylated (hmC/mC). Using single-molecule imaging, Song et al. [224] showed that hmC is found in this configuration ~60% of the time in mouse genomic DNA, and is found in this context across all tissues.
- Runs of up to 3 CpGs are included to check for conversion efficiency in consecutively methylated or hydroxymethylated regions. We were curious whether TET would cope equally well in a region with a lone methylated CpG compared to a cluster of highly methylated CpGs. Identifying technical limitations in these regions would be particularly important for CpG islands.
- Methylated Cs in non-CpG contexts are also included for completeness to supplement the 160mer.
- We included a T(hmC)GA site for restriction-enzyme (TaqI) aided conversion quantification which was done pre-sequencing.
- The edges of the spike-in (the pads) are free of bases of interest (hmC and mC), since the treatment is more likely to degrade free ends.

The resulting spike-in was made of two shorter oligonucleotides (IDT) which anneal across 30bp when combined and are extended with hmC to generate a 250bp double stranded spike-in (Figure 12C). The 80mer containing C, mC, and hmC can be used for conversion rate estimates at HG/HG, HGHG/HGHG, MG/HG, and THGA/THGA sites (H = hmC, M = mC). Providing multiple measures of conversion rate at symmetrically hydroxymethylated sites. The 200mer contains C and mC, and can be used to determine conversion at hemi-hydroxymethylated sites, both in mC/hmC and C/hmC

contexts. The 30bp overlap (filler) contains a single HG/MG site, allowing for estimation at a position that hasn't been experimentally filled.

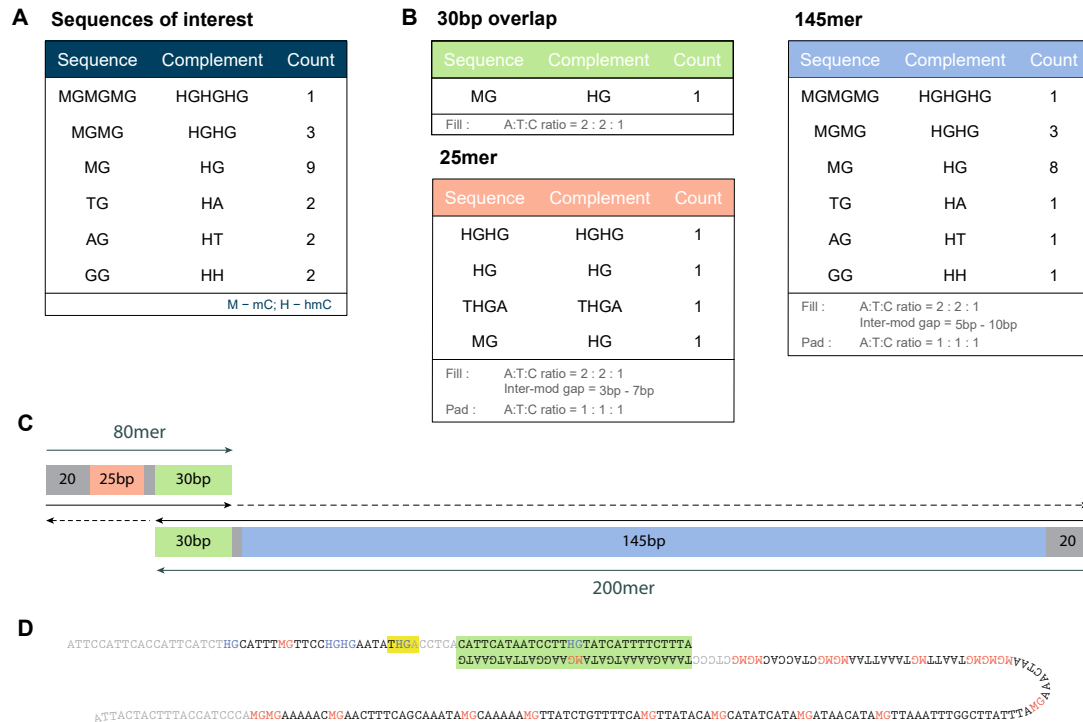


Figure 12: **Spike-in Design.** A. Desired sequences and their frequencies. For example, there will be a site with two adjacent hmC's (HGHG) which will be complementary to mC's (MGMG) so that conversion can be evaluated when there's a run of methylated CpGs. B. The sequences falling into each section of the spike-in. For each portion, the "Sequence"s are inserted, and then the bases are filled according to the ratios stated. The gray pads around the 25mer and 145mer are filled with a slightly different ratio. C. The complete design, with direction of filling with hmC shown. D. The sequences chosen.

As an example, to generate the candidate 145mer spike-ins, a script was written to:

- Randomly sample the order of the 15 sequences (note frequencies in Figure 12B), i.e. generate a permutation of 15 objects, $\{s_1, s_2, \dots, s_{15}\}$. In total, these account for 40bp of the 145bp.
- The 20bp pad is a random combination of A, T, and C in the ratio 1:1:1, similarly for the 5bp pad near the 30bp overlap, $\{p_5, p_{20}\}$ (Figure 12C).

- The remaining 80 bases are split into 14 filler regions of lengths 5-10bp, with A, T, and C in the ratio 2:2:1, $\{f_1, f_2, \dots, f_{14}\}$.
- The individual sequence components are then combined to make the full 145bp spike-in, $\{p_5, s_1, f_1, s_2, f_2, \dots, s_{14}, f_{14}, s_{15}, p_{20}\}$.

An equivalent approach was used to generate candidate 30bp overlap and 25mer sequences. I generated over 100 sequences for each section, combined them to make candidate 250mer's, filtered out any cases with more than one TCGA site, and then mapped each candidate to a merged reference containing hg19, mm10, and any existing spike-ins being used by the labs. In collaboration with Dr. Hiromi Tagoh and Dr. Paulina Siejka-Zielinska we picked the candidates that were most likely to work well²⁹, and chose one from this list to be the 250mer used across all TAPS- β experiments (Figure 12D).

Dr. Paulina Siejka-Zielinska ran both TAPS and TAPS- β for each condition, i.e. unstimulated, +5-aza-CdR, and +Dox, and included the four spike-ins in the experiment (methylated lambda, unmethylated 2.kb, 160mer, and the 250mer designed specifically for this project). This resulted in six samples from which to calculate the modification levels (and changes) after +5-aza-CdR and +Dox treatment.

Spike-in conversion was analysed in these samples to check for conversion rate biases or outliers. The methylated lambda DNA spike-in contains 6,225 unique CpGs, 6,451 CHGs, and 11,503 CHHs. The median coverage per CpG ranged from 86 to 216, allowing for robust estimates of modification level to be calculated. The lambda spike-in showed a mean 95% conversion of mC in a CpG context with little variability between samples (Figure 13). The unmethylated 2kb spike-in (340 CpGs, 290 CHGs, 495 CHH) had median coverage of 122-123, and the mean conversion at CpGs was 0.2%. Thus the mC level at an individual CpG may be under-estimated, but is unlikely to be over-estimated based on these two spike-ins.

The 160mer was used to evaluate context-specific conversion. This spike-in was covered extremely highly in all samples (30k+ reads) as each 150bp read overlapped

²⁹We wanted to avoid secondary structure formation for example.

Modifications	CpG	CHG	CHH	CNN
Methylated lambda spike-in				
5mC+5hmC in unstimulated	95.354	0.235	0.112	0
5mC+5hmC in +5-aza-CdR	95.529	0.225	0.108	0
5mC+5hmC in +Dox	95.524	0.230	0.111	0
5mC in unstimulated	94.764	0.228	0.116	0
5mC in +5-aza-CdR	95.056	0.226	0.109	0
5mC in +Dox	94.683	0.224	0.115	0
Unmethylated 2kb spike-in				
5mC+5hmC in unstimulated	0.232	0.090	0.095	0
5mC+5hmC in +5-aza-CdR	0.190	0.078	0.089	0
5mC+5hmC in +Dox	0.188	0.081	0.086	0
5mC in unstimulated	0.191	0.070	0.134	0
5mC in +5-aza-CdR	0.213	0.068	0.095	0
5mC in +Dox	0.198	0.090	0.068	0

Figure 13: **Conversion rate (%) in fully methylated lambda and unmethylated 2kb spike-ins.** The conversion rate per CpG is calculated as $\text{MOD}/(\text{MOD} + \text{UNMOD})$. Table shows context-specific mean conversion rates per sample (H - any of A, C, T).

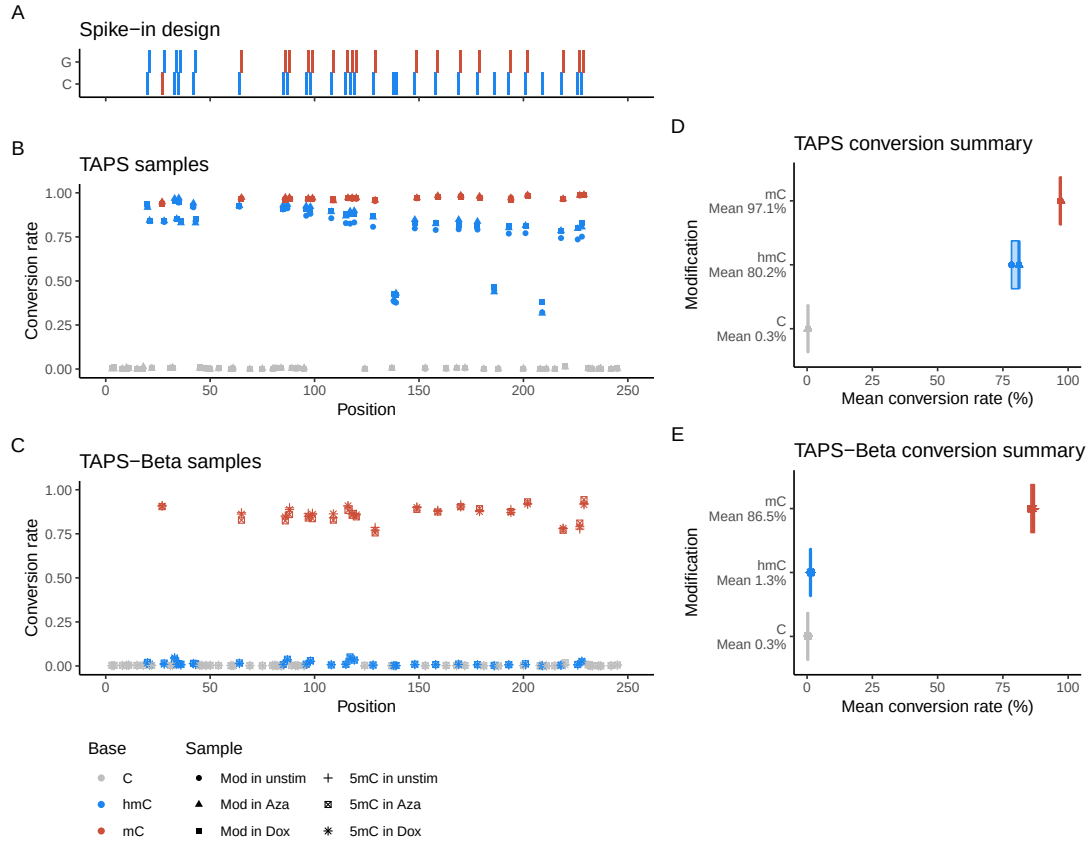


Figure 14: **Conversion in the 250mer.** A. The design of the spike-in, with hmC positions shown in blue and mC positions in red. B-C. Per-base conversion rate in TAPS and TAPS- β samples, respectively. D-E. Boxplots showing the mean conversion rate across all modified positions in TAPS and TAPS- β samples, with the mean across all samples annotated.

the two positions of interest. The contexts had a wide range of coverage but the minimum (73 reads) was more than sufficient to calculate context-specific conversion. CpG sites had clearly higher conversion efficiency compared to CpH and the protection from conversion at hmC in the TAPS- β samples was confirmed (Figure 15). The rate of conversion outside CpGs is clearly variable, with C's preceded by T's the least efficiently converted. However, since the majority of methylated positions in humans are found in a CpG context, our analyses focused on these only³⁰.

To get further insights into hmC conversion, we analysed the 250mer (the positions of each mC and hmC are shown in Figure 14A for reference). The per-base conversion in the TAPS sample was better at mC compared to hmC with mean conversion of 97.1% and 80.2%, respectively, and showed only 0.3% conversion at unmodified CpGs (Figures 14B and D) with little position-specific variation. This 80.2% value is biased by the hmC sites which were included in a non-CpG context (positions 138, 139, 186, 209) and without these the mean conversion rate is 86%. The conversion rate is also worse near the 3' end of the spike-in, but there was no such trend for mC in this set of samples, or other TAPS- β samples that we've analysed.

The conversion in the TAPS- β samples confirms that the hmC positions were largely protected from conversion³¹, meaning that mostly mCs contribute to the modification signal (Figure 14C and E). mC conversion in TAPS- β samples based on this spike-in is worse at 86.5% (compared to 97.1% in TAPS) however since most mC in the genome is found in an mC/mC context, the fully methylated lambda spike-in is a more relevant measure of conversion for mC.

Based on the positions of highest hmC conversion, it's clear that the unprotected sites seem to lie inside CpG clusters. Indeed the sites with highest hmC conversion are 117 (4.5%), 33 (4.3%), 87 (3.6%), and 119 (3.4%), which are all surrounded by mC on the opposite strand. Since the ultimate goal is to profile mC and hmC across the human genome, we calculated the frequencies of lone CpGs and CpG runs to ascertain how much effect this reduction in hmC protection has in the context of the entire genome.

³⁰Based on these results TAPS and TAPS- β should not be used when evaluating non-CpG methylation, such as CpA.

³¹Although not negligible at 1.3%, conversion is sufficiently low to consider most hmC sites protected.

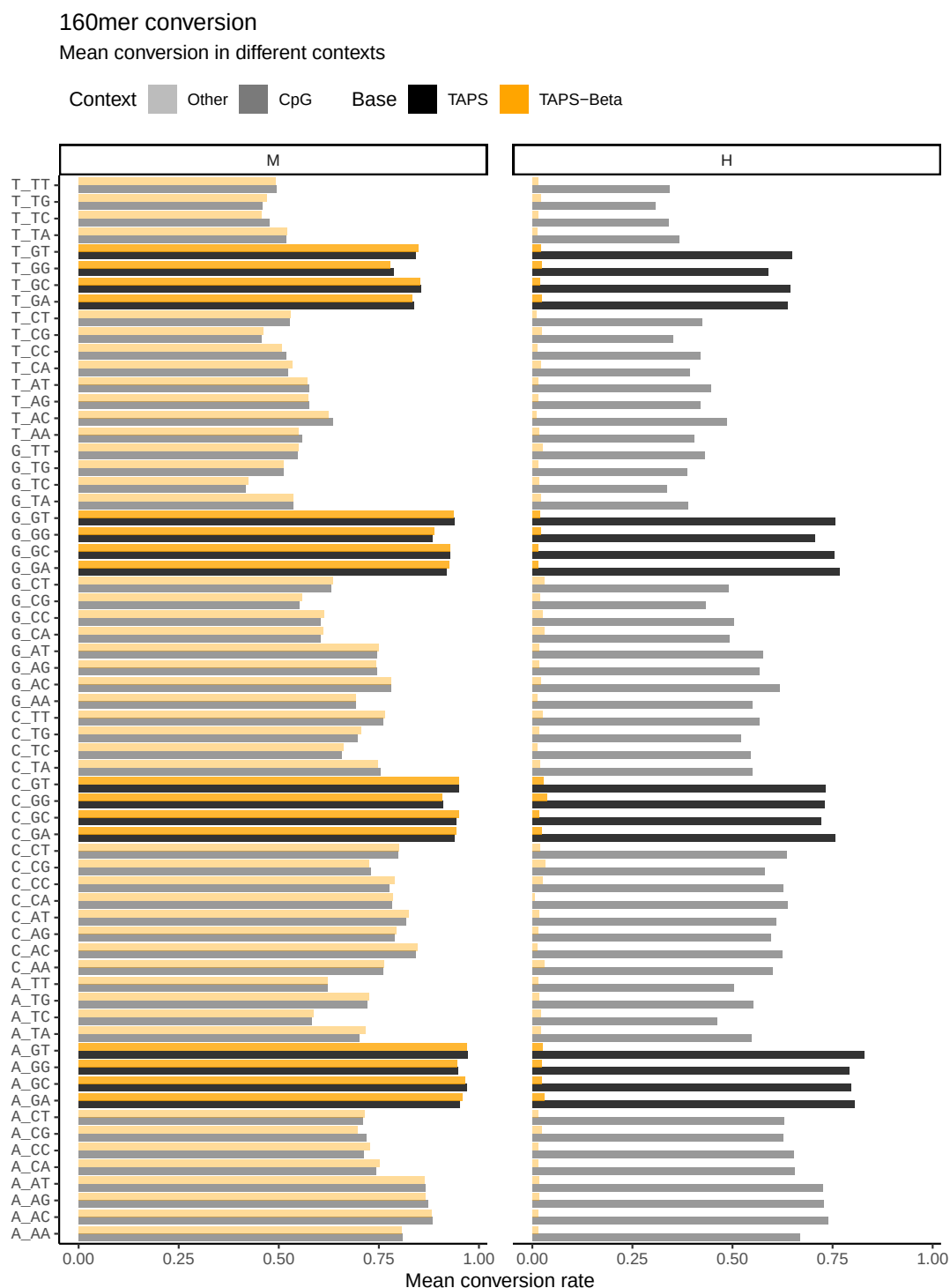


Figure 15: **Conversion in the 160mer.** Context-specific conversion rate in TAPS and TAPS- β , where M is mC and H is hmC. The conversion rates for each context-experiment pair are summarised across the three conditions.

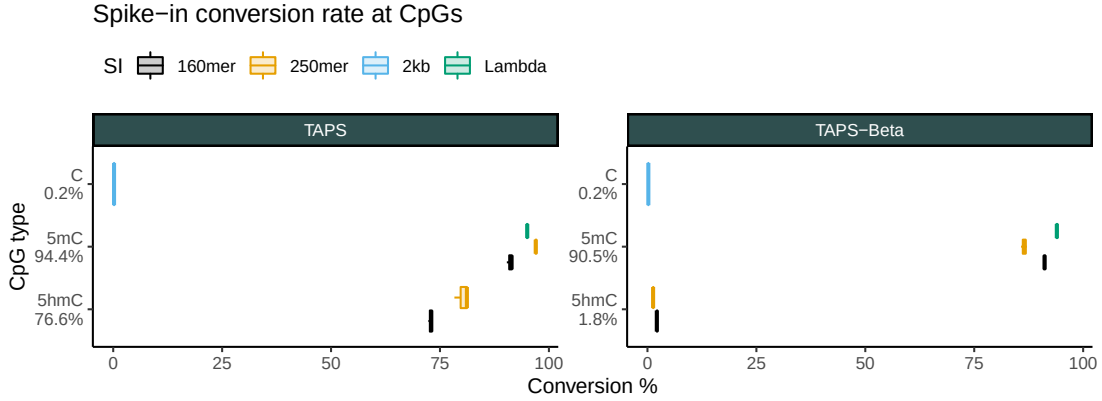


Figure 16: **Conversion compared across spike-ins.** Boxplots showing CpG conversion summarised per sample for each spike-in.

Lone CpGs account for the vast majority (92.8%) of CpGs in hg19³² (Table 3), therefore since CpG runs are rare our results are unlikely to be biased by this phenomenon.

Pattern	#CpG	Count	% CpGs
E CG F	1	26,183,412	92.8%
E CGCG F	2	593,823	4.21%
E CGCGCG F	3	36,248	0.385%
E CGCGCGCG F	4	4,941	0.07%
E CGCGCGCGCG F	5	1,461	0.026%
E CGCGCGCGCGCG F	6	685	0.015%

Table 3: **CpG run count.** Frequency of Pattern in the hg19 genome (chromosomes 1-22,X,Y only), where E = i) A, C or T, or ii) G preceded by A, G, or T, and F = i) A, G, or T, or ii) C followed by A, C, or T. Calculations carried out using `stringr::str_count` in R.

To summarise, the data from the spike-ins shows that the TAPS and TAPS- β experiments resulted in six samples with highly comparable conversion i.e. conversion biases are unlikely to affect our results. The conversion at mC was sufficiently high, meaning we can accurately detect methylation at base resolution both before and after +5-aza-CdR and +Dox treatment without much contamination from false positives (Figure 16). Based on the 160mer and 250mer, conversion at hmC is worse, suggesting that underestimates of hmC level of over 20% are likely, but since we will ultimately be looking for changes in mC and hmC level, as long as the underestimation is consistent we can still interpret the data.

³²Chromosomes 1-22, X, Y.

3.2.2.1.2 Coverage

To evaluate how many CpGs we can confidently assign mC and hmC levels to, we calculated the coverage at each CpG. The TAPS and TAPS- β samples covered (on average) over 97% of CpGs with 1+ reads³³, but since lowly covered positions can result in greater errors in modification calculation, we filtered the CpGs using a minimum coverage cut-off. To choose a sensible threshold, we used the coverage distribution for each sample (Figure 17) and found that the unstimulated TAPS- β sample had the lowest coverage and thus would be the most affected by the magnitude of the cut-off chosen. Since this sample would be used to calculate hmC in unstimulated, mC gain after both treatments, and hmC gain after both treatments, a less stringent threshold of 5+ reads per CpG was chosen (rather than 10+) when defining a CpG as sufficiently covered for further analysis. This resulted in a mean of 51,894,932 covered CpGs per sample, or 92% of total CpGs after filtering (Figure 18A).

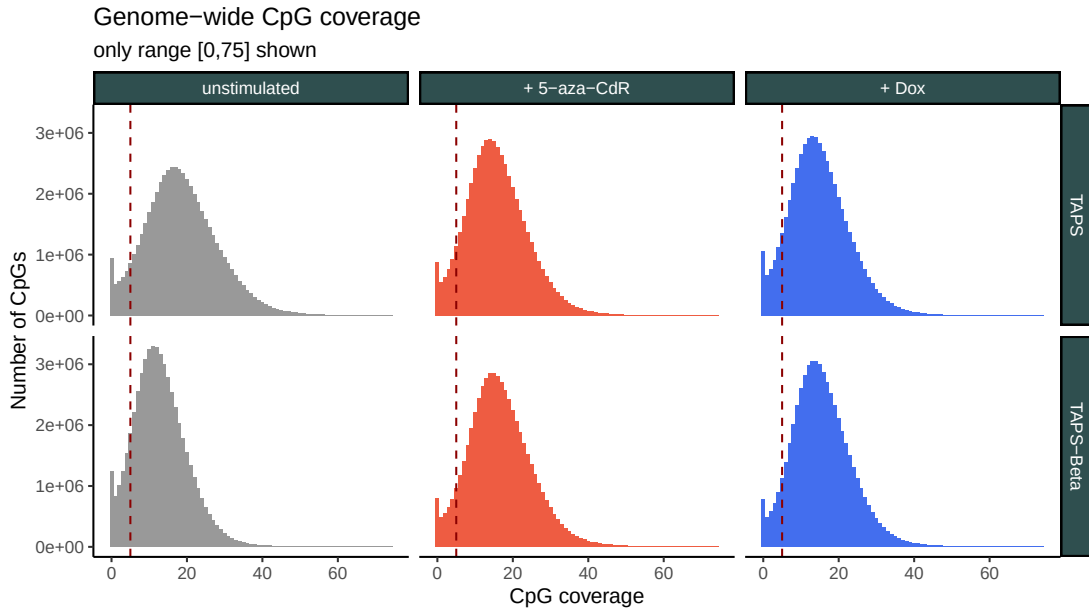


Figure 17: **CpG coverage.** The densities of coverage show that some samples, in particular unstimulated TAPS- β , have fewer highly covered CpGs.

Recall that we want to calculate the gains and losses in mC and hmC upon +5-aza-

³³Chromosomes 1-22,X,Y.

CdR and +Dox treatment. To this end, the TAPS and TAPS- β modification levels will need to be combined and contrasted per position, e.g. to calculate mC loss at position X after +5-aza-CdR we must compare TAPS- β unstimulated with TAPS- β +5-aza-CdR which requires both samples to have sufficient coverage at X . After using a coverage threshold of 5+, we only have hmC gain measurements at $\sim 86\%$ of CpGs, but the other modification gains and losses can be calculated at almost 90% of CpGs (Figure 18B).

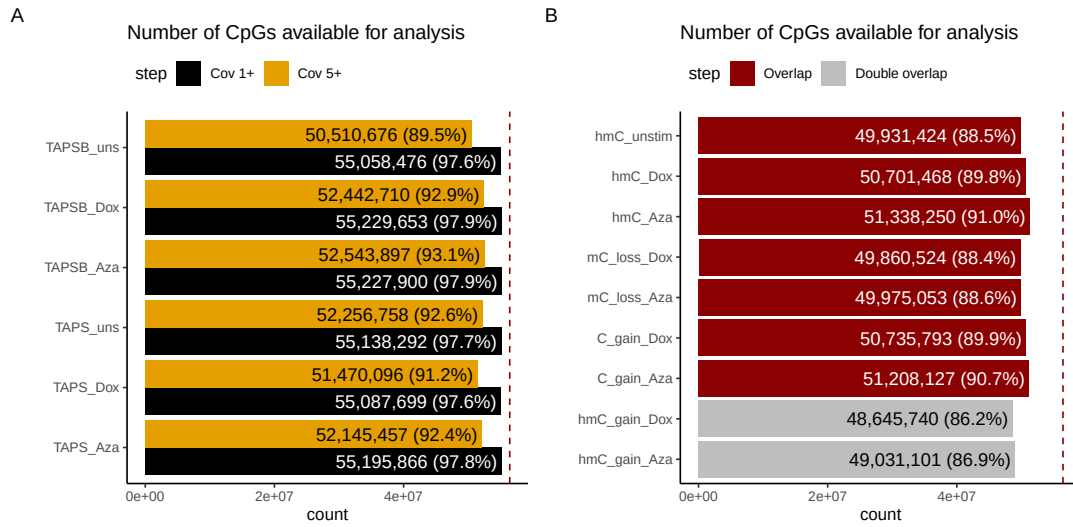


Figure 18: **Number of CpGs after each step.** A. Bar plots showing the number and percentage of CpGs before and after filtering out CpGs with coverage 1-4. B. The number of CpGs with sufficient coverage to calculate hmC level, mC loss, hmC gain, and unmodified C gain.

3.2.2.1.2.1 Copy Number Analysis

Another aspect to consider when picking a coverage threshold is copy number. HEK293 is a hypotriploid human cell line with multiple duplications, deletions, and varying karyotype [145, 227]. Although we did not expect differences in karyotype between conditions³⁴, picking a threshold that would eliminate certain regions from the analysis due to deletions was a worry.

³⁴Each experiment was performed on cells in the same culture, split into conditions and treated (or not) for 48 hours, and any significant karyotype alterations are very unlikely within this 48 hours.

We used coverage data to check for copy-number changes. Based on the mean coverage across 10kb bins from the TAPS unstimulated sample, there were some clear copy number gains in multiple chromosomes with some evidence of losses (chr19), meaning that some CpGs are expected to have higher (or lower) coverage conditional on karyotype (Figure 19). However, since we're using a substantially low threshold of 5+ reads, and there is indeed no copy number differences induced by the treatments based on the profiles of the +5-aza-CdR and +Dox cells looking very similar (Figure 20), we expect this copy number variation to have minimal effect on the following analyses.

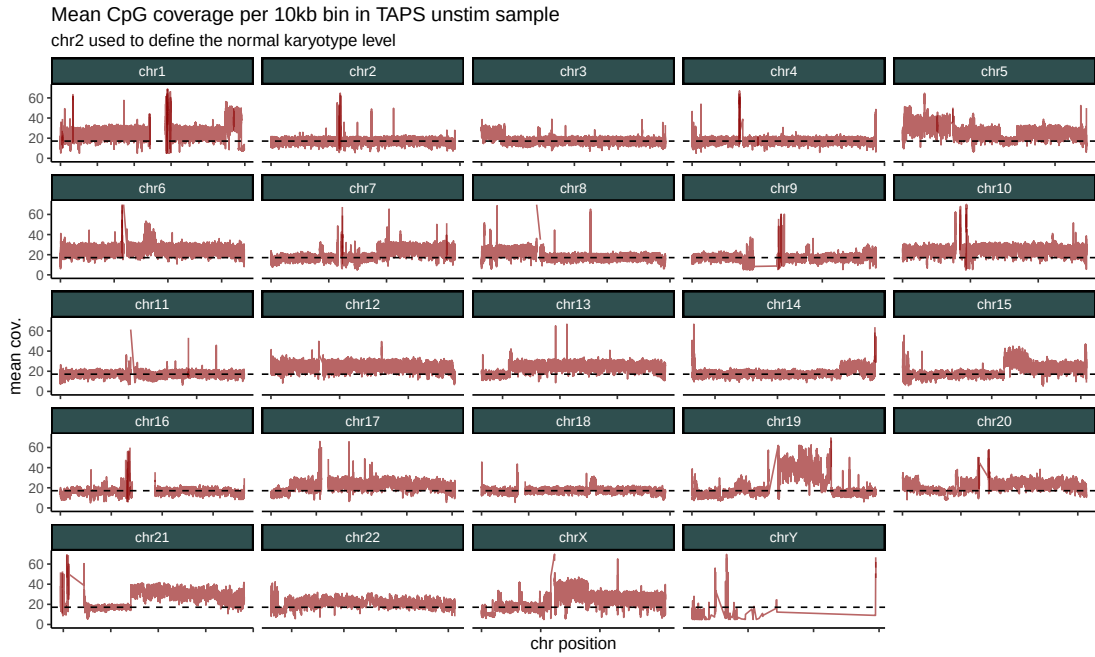


Figure 19: **CpG coverage in 10kb bins to evaluate copy number differences.** The mean coverage across 10kb bins shows clear regions of duplication in the untreated cells.

3.2.3 Genome-wide base-level analysis of induced modification changes

The aim of the experiment was to achieve *unbiased* demethylation with +5-aza-CdR, and *unbiased* oxidation by inducing the TET1-CD with Dox.

The distribution of modification levels shows that the untreated sample had little hmC (since the densities of TAPS and TAPS- β are very similar), and most positions

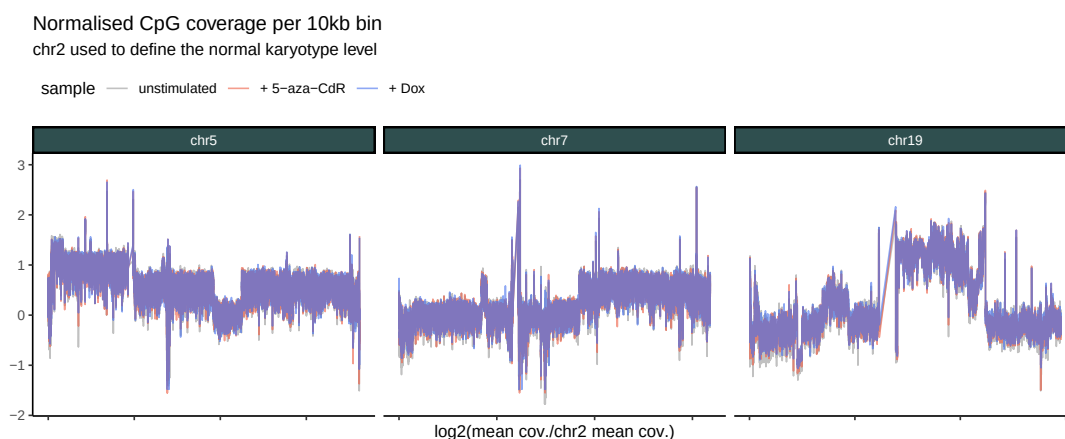


Figure 20: **Normalised CpG coverage in 10kb bins to compare copy number between conditions.** The mean coverage across 10kb bins in each sample is normalised by the mean coverage of chromosome 2. A subset of the chromosomes with clear copy number gains and losses was chosen for illustration.

were either unmethylated (mod. level = 0) or highly methylated (mod. level ~ 1) (Figures 21A and B). Supposing the +5-aza-CdR treatment targeted a specific subset of modified CpGs, then, based on the unstimulated analysis these would begin as fully methylated. We would then expect to shift those CpGs from the right (fully modified) to the left (unmodified) peak of the unstimulated distribution, without increasing the proportion of CpGs with intermediate modification levels substantially. However, in this case all the CpGs appear to lose methylation uniformly and therefore +5-aza-CdR doesn't appear to have targeted effects.

The +Dox treated sample shows the most difference between TAPS and TAPS- β , meaning a significant proportion of the modified sites are hmC. The TAPS- β curves for +Dox and +5-aza-CdR suggest similar levels of mC reduction (which is consistent with the MS results).

3.2.3.1 Calculating hmC level and changes in mC

In Section 3.2.1.3 we selected suitable concentrations of +Dox and +5-aza-CdR based on mass spectrometry data to achieve similar mC loss after both treatments while the +Dox sample also gained hmC. Here we would like to confirm, at base resolution, using

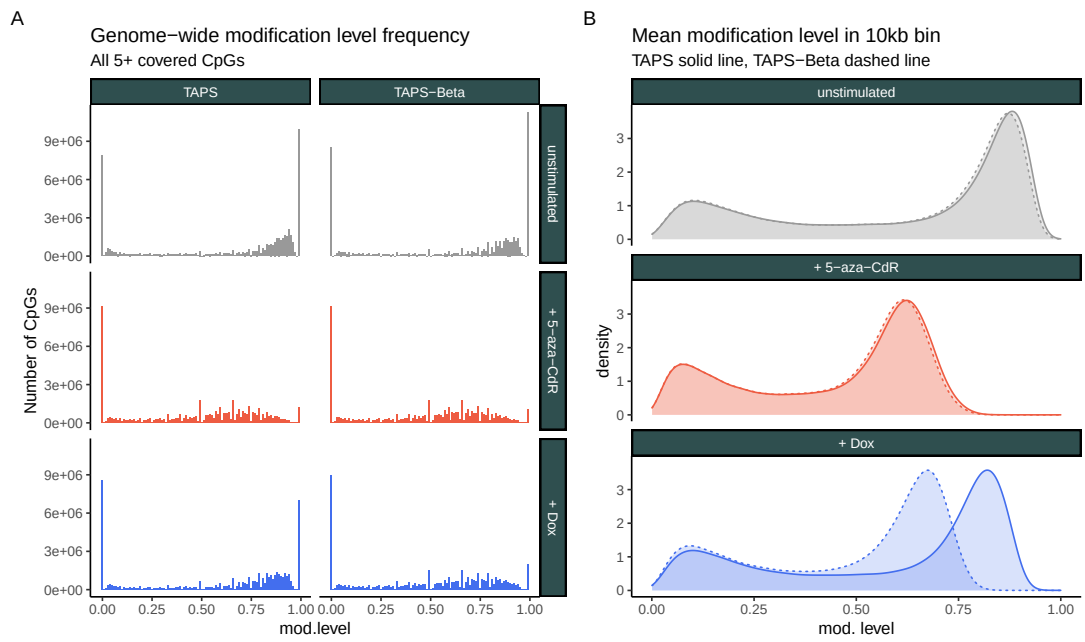


Figure 21: **Modification levels compared.** A. Frequencies of modification level in each sample. B. The densities of mean modification level in 10kb bins across the genome. In each case, only CpGs with 5+ reads were considered for the analysis.

an orthogonal approach that we were successful in achieving these genome-wide changes.

TAPS (mC+hmC) and TAPS- β (mC) data were used to determine the level of hmC in each condition by subtracting the conversion rate at each CpG. Genome-wide, most CpGs have hmC level close to zero, suggesting that the position only showed evidence of mC (Figure 22A). In contrast, the +Dox treated sample shows a clear tail in the hmC calculation which confirms substantial hmC gain. Noting that our confidence in the conversion rate differences observed is dependent on the coverage, we used a one-sided proportion test to calculate a p-value associated with each hmC level to confirm that the hmC level after TET1-CD induction was indeed substantially higher³⁵ (Figure 22B).

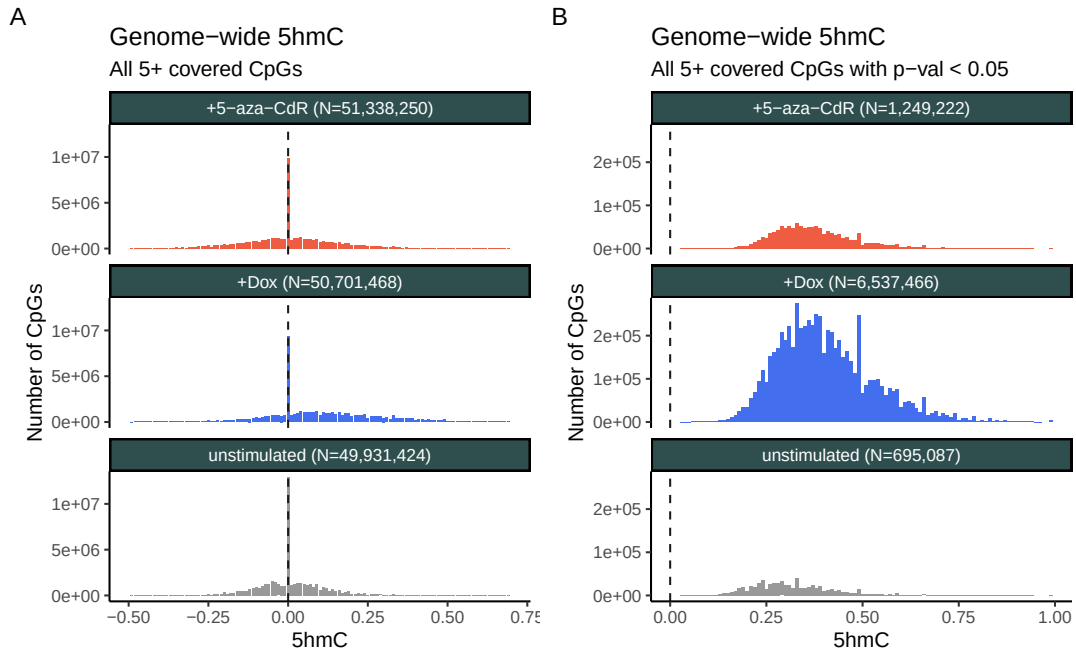


Figure 22: **Genome-wide hmC level.** A. Frequencies of hmC proportion per CpG, calculated as TAPS - TAPS- β , summarised across all 5+ covered CpGs. B. Frequencies of hmC level at CpGs that show significant (p-val < 0.05) changes based on proportion test.

By comparing the mC and mC+hmC levels between unstimulated and treated we calculated the mC loss and unmodified C (uC) gain³⁶ at each CpG, respectively. mC

³⁵See Section 2.2.

³⁶uC gain is equivalent (mC + hmC) loss, and can be thought of as gain in modification free

loss was slightly more extensive in +5-aza-CdR compared to +Dox when looking at each CpG separately, and the slight discrepancy between the two can be seen in the MS result (Figures 11 and 23A). We used the same approach as above to assign p-values to changes at each position for each comparison and found clear differences between the mC loss and C gain in +5-aza-CdR (which should be equivalent given very little hmC). The differences can be explained by the fact that the conversion rates in the TAPS and TAPS- β samples differed (mC loss is calculated based on a $\sim 90\%$ converted sample whereas C gain on a $\sim 95\%$ converted sample).

Similarly, the negative mC loss and uC gain values represent positions that *gained* methylation upon treatment, and since both treatments reduce methylation, this may be the result of secondary effects or a consequence of incomplete conversion during TAPS and TAPS- β . This, combined with the negative hmC levels encouraged us to look more closely at the effects of incomplete conversion.

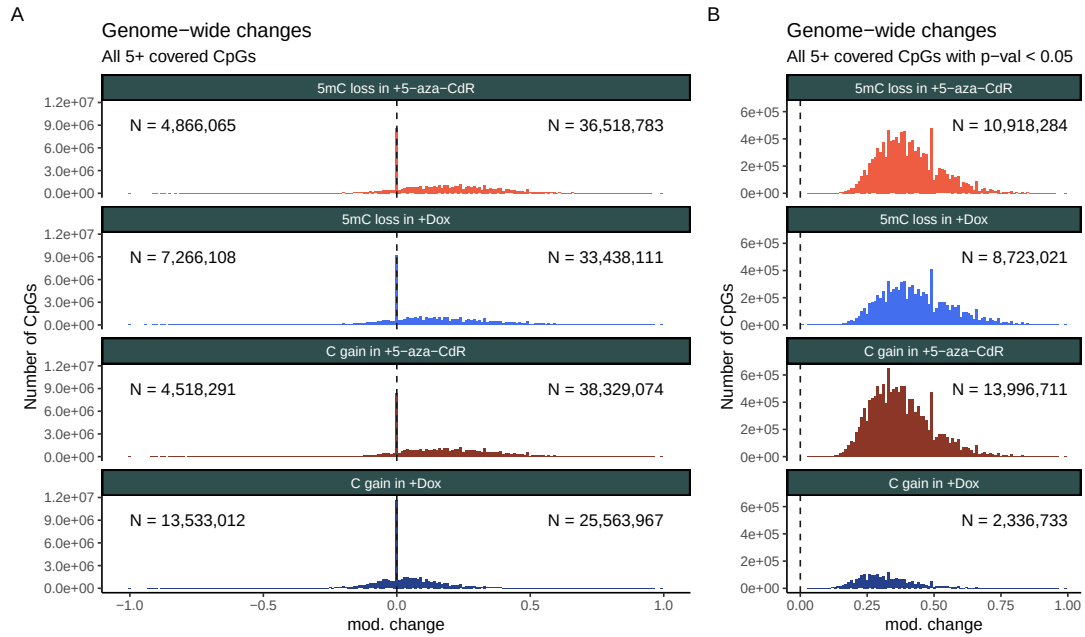


Figure 23: **Genome-wide changes.** A. Frequencies of modification level differences per CpG, summarised across all 5+ covered CpGs, based on TAPS and TAPS- β (see Table 1). B. Frequencies of modification level differences that are significant based on proportion test (p-val < 0.05).

CpGs.

3.2.3.2 The effects of incomplete conversion

Conversion rate provides an estimate of the true modification level at a position, and the accuracy of that estimate depends on the efficiency of conversion and coverage. It is clear, based on the spike-ins, that the conversion efficiency isn't 100% and differs between mC and hmC, therefore our estimated modification levels will be more likely to underestimate modification level but also have some stochastic variation. To check whether the *gains* in mC and negative hmC proportions can be explained by this, we analyse such incomplete conversion through simulation.

We simulated binomial distributions with varying success probabilities (conversion rate) and trial numbers (coverage) (Figures 24A and B). In the context of modifications, we have generated modification level values for 10,000 fully methylated CpGs subject to incomplete conversion in two samples, and then subtracted the modification levels to calculate the difference (which should be zero as both samples were fully methylated to begin with). Both conversion rate and coverage affect the width of the distribution.

In most methylation analyses where the modification level is rounded to 0 or 1 based on some threshold, 90% conversion would result in an average of 10% loss in conversion at each position. Assuming a threshold of 0.5, positions with a true modification level of $0.5 - 0.55$ could be incorrectly considered as unmethylated, but with a majority of true modification levels at 0 or 1, this affects a very small proportion of the data. In our data however, we expect partial demethylation ($\sim 20\%$) at each position, therefore using binary methylation level in the treated samples would mask a significant proportion of the changes.

To check how well our data aligns with the above explanation, we simulated TAPS and TAPS- β samples using the coverage frequencies from the data, a binary modification level (70% of positions were modified), and mean conversion levels calculated from the spike-ins (Section 2.2). The results of the simulations are strikingly similar to the data obtained (Figure 25). The hmC levels calculated in the simulated unstimulated sample, assuming minimal hmC content per position (modelled by an exponential distribution with mean 0.0001, Figure 25B), showed the same expected peak at zero with tails reaching in both directions (Figure 25A), suggesting that on average these tails should

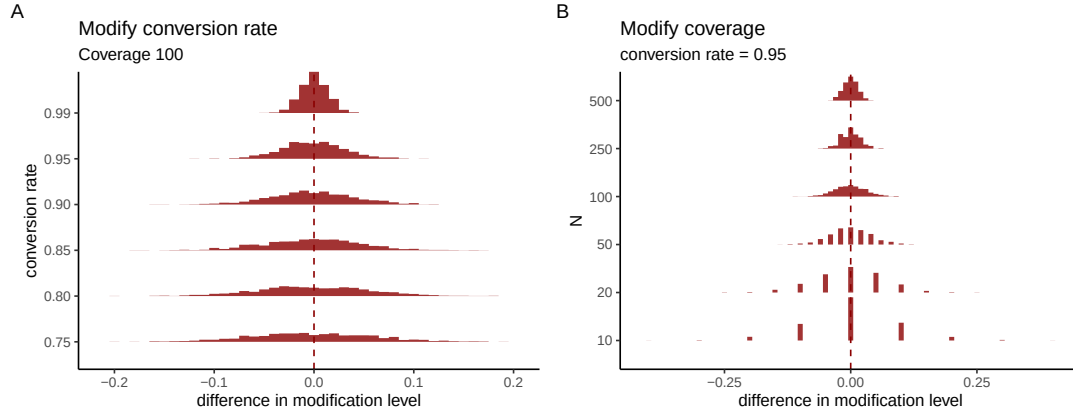


Figure 24: **Behaviour of binomial distribution.** Frequency counts of modification level differences between two fully modified samples with 10,000 CpGs subject to different conversion rate (A.) and coverage (B.).

be considered as stochastic noise in positions with low coverage (conversion rate is fixed).

To show the expected distribution of random, genome-wide demethylation, we simulated another unstimulated TAPS- β sample (with 80% methylation since this had a stronger resemblance to the data in 21A) and generated a +5-aza-CdR sample with demethylation proportion $x \in \max(0, N(\mu = 0.2, \sigma = 0.2))$ per CpG (Figures 25C and D). The distribution of the simulated modification level after demethylation in Figure 25E resembles the shape of levels we calculated in Figure 21 (the peak can be widened and shifted by modifying the demethylation distribution). Notably, we see evidence of negative mC loss (i.e. mC gain) despite true levels always showing loss (or no change).

These simulations support the hypothesis that false negatives are adding noise to the results. Binarising the treated samples would mask a lot of changes, however, based on the modification level distributions in the *unstimulated* TAPS and TAPS- β samples, should we binarise these? Noting that since the conversion efficiency is the same across the samples within each protocol and we are most interested in relative differences between samples (gains and losses), forced binarisation of a subset of samples would over-estimate the modification changes in the +5-aza-CdR and +Dox samples³⁷.

³⁷Since we'd be comparing an estimated 100% converted unstimulated sample with a 90% converted sample, for example.

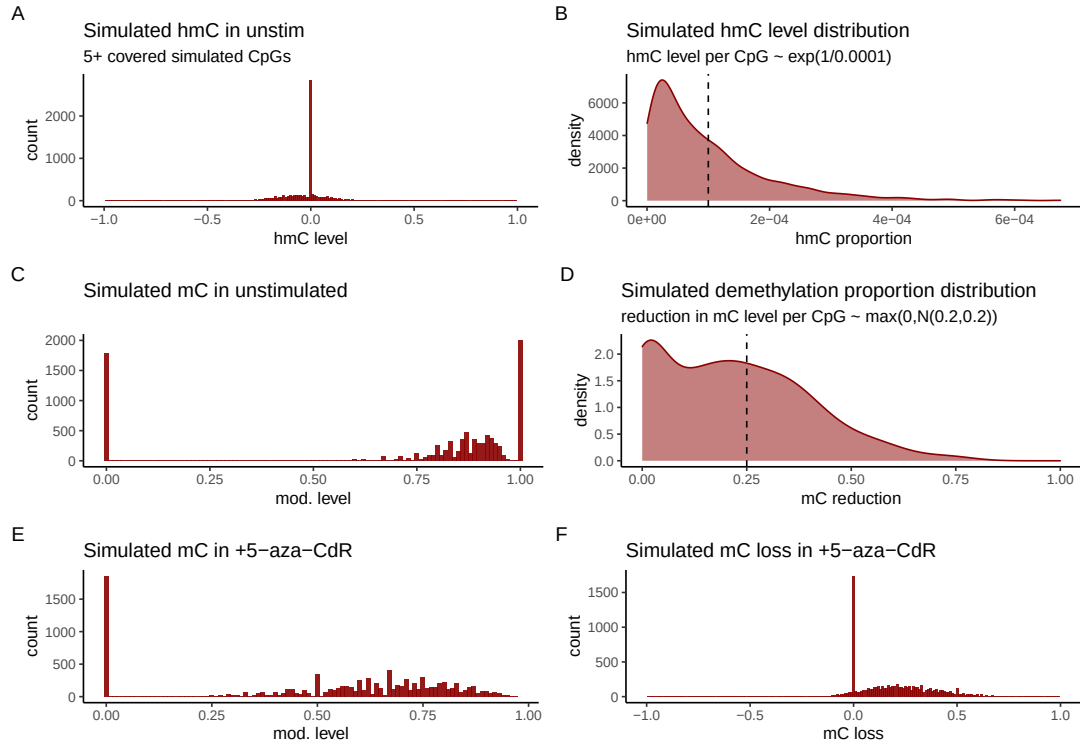


Figure 25: **Simulations of conversion.** A. Distribution of simulated hmC level in the unstimulated sample. B. Density of simulated hmC level per CpG based on an exponential distribution with mean 0.0001. C. Distribution of simulated mC level in unstimulated. D. Simulated reduction in mC based on a normal distribution, with mean 0.25 and standard deviation 0.1. Since a reduction cannot be negative, any values below zero were rounded up to zero. E. The mC level after simulating +5-aza-CdR treatment combined with conversion efficiency. F. Frequency of changes in mC when comparing simulated mC levels in unstimulated and +5-aza-CdR.)

We also used this simulated data to evaluate the efficacy of using the proportion test to pick out the most affected CpGs (Figure 26). Many of the true changes were lost through filtering, most over-exaggerating the mC level post-conversion compared to the true effect. Therefore we chose to instead use averages of the unfiltered values (relying on the Central Limit Theorem) in our analyses and avoided looking at individual CpGs.

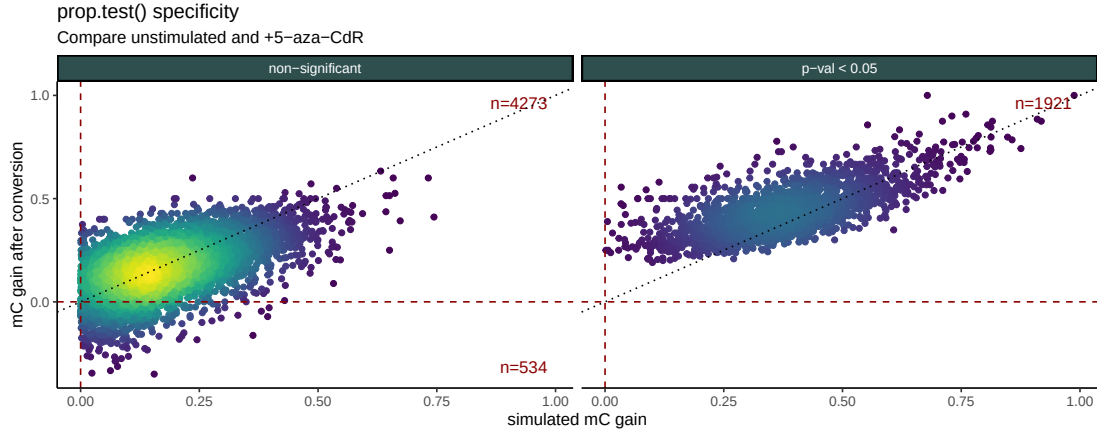


Figure 26: **Effect of checking for significant changes.** Density scatter plots comparing true simulated mC gain and mC gain after incomplete conversion. Each point represents a CpG with 5+ coverage in both simulated samples, which are split according to the calculated one-sided proportion test p-value (based on the post-conversion modification count and coverage). The number of CpGs in each quadrant are shown in red.

3.2.4 Genome-wide analysis of trends in induced modification changes

We calculated the mean modification levels in 10kb, non-overlapping bins to look at changes genome-wide. This resulted in a median of 145-149 5+ covered CpGs per bin (both strands included) which are combined. If the changes are indeed stochastic, then the means should tend to the true mean of each bin.

As expected, hypermethylated regions (mC in unstimulated > 0.75) have the most potential to lose methylation (Figures 27A and B). However, we also note the discrepancy between the treatments in that +5-aza-CdR mC loss appears more linearly related to mC in unstimulated, whereas the TET1-CD upregulation acts less on regions with intermediate baseline methylation, i.e. mC unstim. $\in [0.25, 0.75]$. The direct compari-

son of mC loss confirms that +5-aza-CdR is more consistent at demethylating compared to +Dox, with fewer bins that showing no change in mC level i.e. $\text{mC loss} = 0$ (Figures 27C and D). To investigate whether this phenomenon may be a consequence of accessibility, we calculated the inter-CpG distances on chr19, and found the maximum, minimum, and mean distance between closest neighbours for each CpG (on the same strand) (Figure 28A). Summarising these quantities over 40k randomly sampled CpGs in each methylation group:

- Unmethylated: $\text{mod. level} < 0.3$
- Intermediate: $0.3 \leq \text{mod. level} \leq 0.7$
- Methylated: $0.7 < \text{mod. level}$

we found that CpGs with intermediate mC levels were more likely to be further from other CpGs (Figure 28B) and may therefore be less accessible. However, since intermediate baseline mC levels are rare, we do not foresee this to be a significant problem.

Are gains in hmC similarly related to baseline mC? In order to see gains in hmC, a position should first be methylated and then oxidised by TET. Therefore, we expect a similarly strong relationship between hmC gain and baseline mC, and indeed the majority of hmC gains were seen in highly methylated regions (Figure 29A). We also confirmed that hmC gain at a position is proportional to mC loss, and found that, on average, hmC gain explains $\sim 71\%$ of the mC loss after +Dox treatment (the remainder explained by loss of both mC and hmC, i.e. unmodified C (uC) gain) (Figure 29B).

3.2.5 Analysis of DNA modifications in functional regions

3.2.5.1 DNA modification changes in genes

With a goal of linking DNA modifications to expression, we wanted to confirm that methylation in genes was indeed affected by the treatments.

Previous methylation maps have shown an accepted mC profile for genes, consisting of a dip at the TSS which corresponds to hypomethylation at promoters, and a subsequent increase in methylation across the gene body which is slightly elevated compared

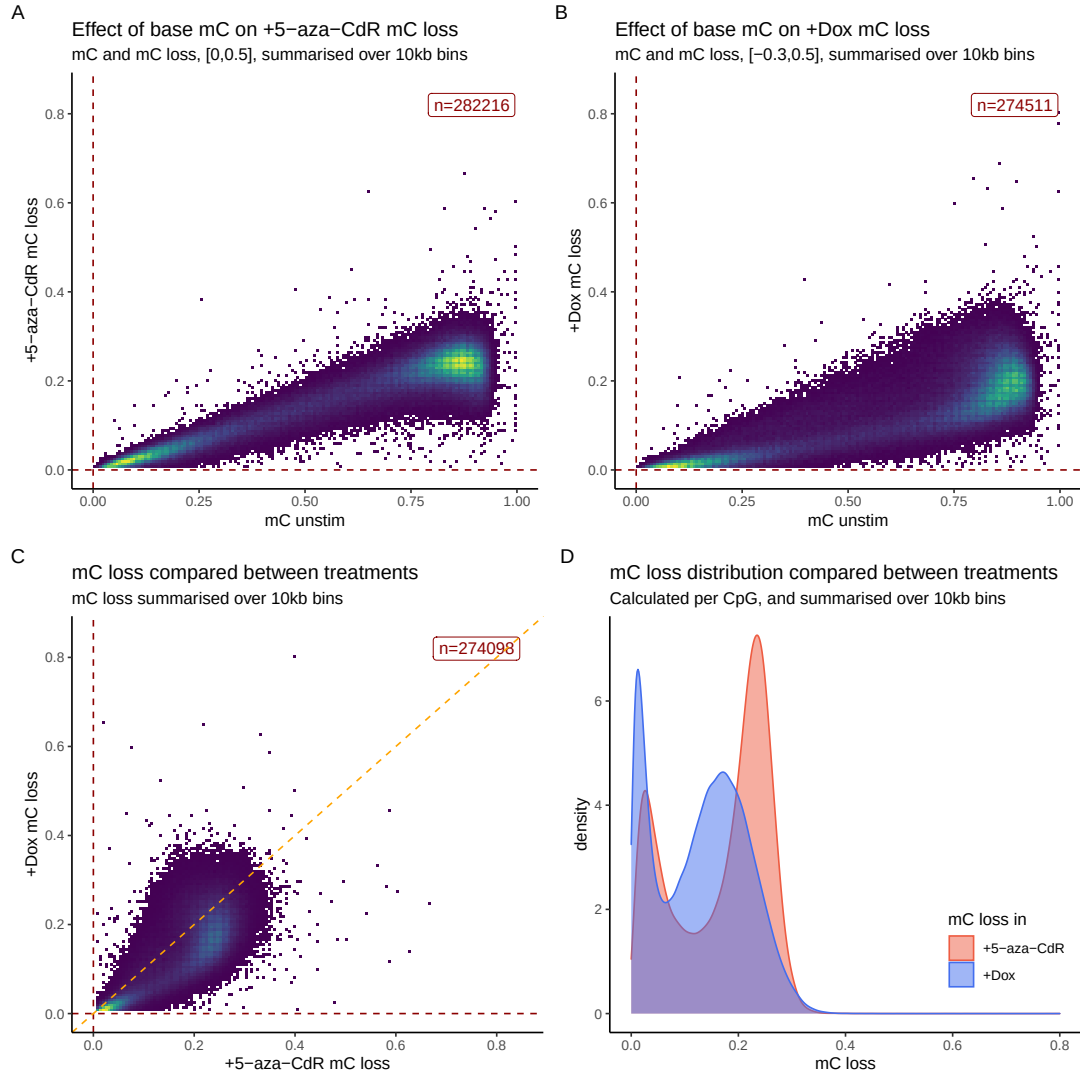


Figure 27: **mC loss in 10kb bins**. A-B. Binned density scatter plots comparing the mean level of mC in unstimulated, calculated per CpG and summarised genome-wide in 10kb, non-overlapping bins and the corresponding mC loss after +5-aza-CdR and +Dox treatments, respectively. C. Direct comparison of mC loss per 10kb bin. D. Densities of 10kb bin mC loss.

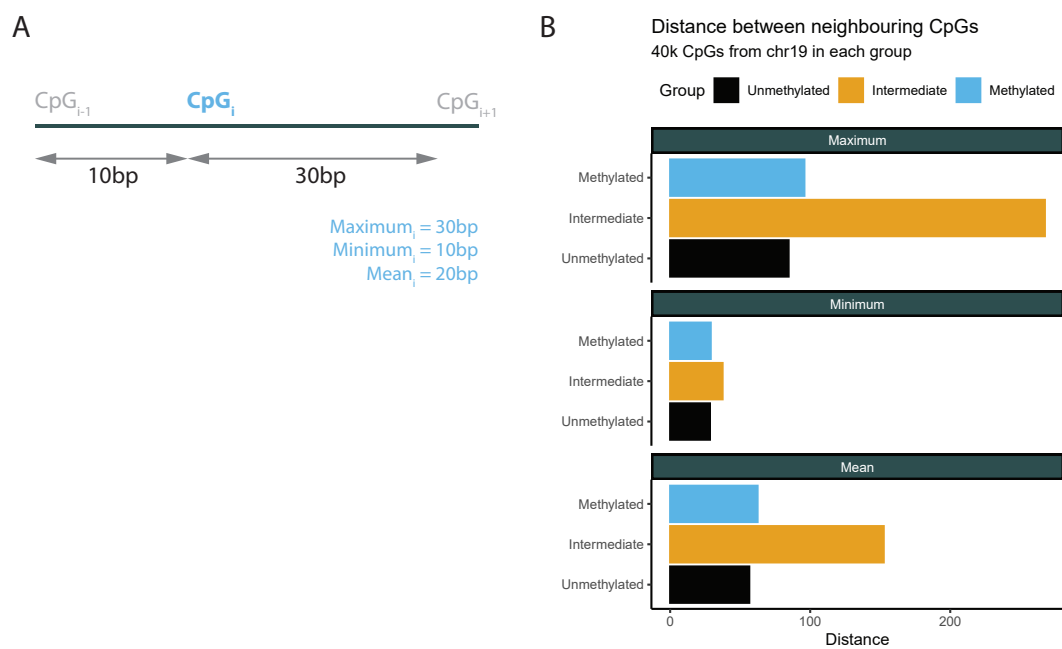


Figure 28: **Distance between neighbouring CpGs.** A. Schematic showing what each of the distances represent. B. Mean of the distances across 40k randomly sampled CpGs from each methylation group on chr19.

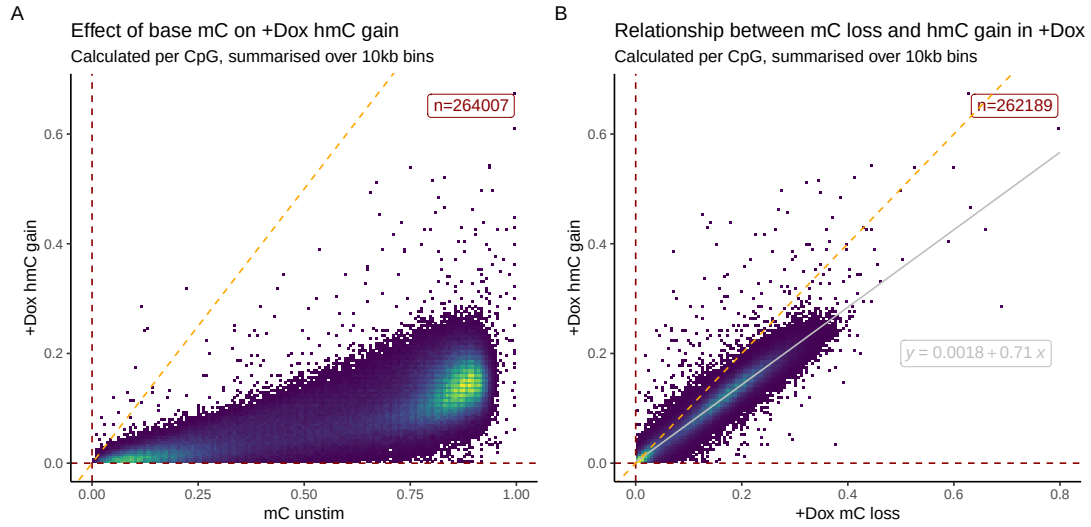


Figure 29: **hmC gain in 10kb bins.** A-B. Binned density scatter plots comparing the mean level of mC (TAPS- β) in unstimulated, calculated per CpG and summarised genome-wide in 10kb, non-overlapping bins and the corresponding mC loss after +5-aza-CdR and +Dox treatments, respectively. C. Direct comparison of mC loss per 10kb bin. D. Densities of 10kb bin mC loss.

to the flanking regions [120, 273, 65]. The mC calculated using TAPS- β in unstimulated HEK293 cells showed a very similar profile (Figure 30A), meaning that gene bodies were more susceptible to methylation changes compared to promoters (for example).

The mC loss in genes after +5-aza-CdR and +Dox treatments is comparable, but with a marginally higher level of mC loss after +5-aza-CdR (Figure 30A). The difference in hmC gain between the two treatments is significant (Figure 30B), and consistent across the genome, with the exception of the TSS, which had little mC to begin with.

Looking closer at the TSS and TES profiles in Figure 31, there appears to be a gradual increase in mC downstream of the promoter. Note that there is a lack of dip at the TES (Figure 31B), which appears prevalent in other studies and in our own plot (Figure 30A). This can be explained in two ways, the percentile-based visualisation often used incorrectly suggests differences in methylation upstream and downstream of the TES due to different bin sizes flanking the TES. Separately, due to the comparably low CpG density in this region (Figures 31E and 31F) we expect that enrichment-based techniques would simply under-estimate methylation in this area due to reduced binding.

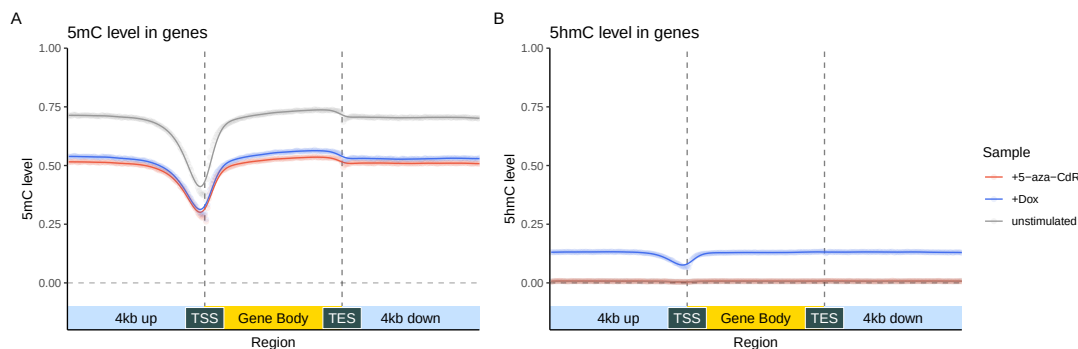


Figure 30: **DNA modifications in genes.** The metagene plots summarise modification levels in the vicinity of genes, with a line of best fit joining the individual (low opacity) means per region for each condition. A. The mC levels calculated based on the TAPS- β data are summarised into bins across gene bodies and adjacent 4kb upstream and downstream regions. B. The hmC levels per CpG are calculated as the difference between TAPS and TAPS- β modification levels, and summarised in the same bins.

3.2.5.2 DNA modification changes in CpG islands

Although CpG islands (CpGIs) are generally unmethylated, we wanted to see whether there were any hypermethylated candidates that lost mC (or gained hmC) upon +5-aza-CdR treatment or TET1-CD upregulation. Therefore we calculated the mean mC and hmC levels in CpGIs as well as their flanking regions, i.e. the shores and shelves³⁸ (Figures 32A and B).

Most CpGIs were unmethylated, but, of the 26,090 CpGIs with at least one covered CpG, 32% had a mean methylation across bins > 0.8 (Figure 32C). On average, these CpGIs appeared hypermethylated compared to the corresponding flanking regions (Figure 32D).

Although the CpG density in CpGIs is high and uniform across the region (Figure 32E), there didn't appear to be a decrease in mC across the region in +5-aza-CdR and +Dox compared to the unstimulated sample (Figure 32A) which seemed strange based on the summarised data (Figure 32C). This suggested a discrepancy between using the mean mC level as a predictor of CpGI methylation and the more granular,

³⁸Given a CpGI with start and end coordinates $[C_{start}, C_{end}]$, we define the shores as $[C_{start} - 2kb, C_{start}]$ and $[C_{end}, C_{end} + 2kb]$, and shelves as $[C_{start} - 4kb, C_{start} - 2kb]$ and $[C_{end} + 2kb, C_{end} + 4kb]$.

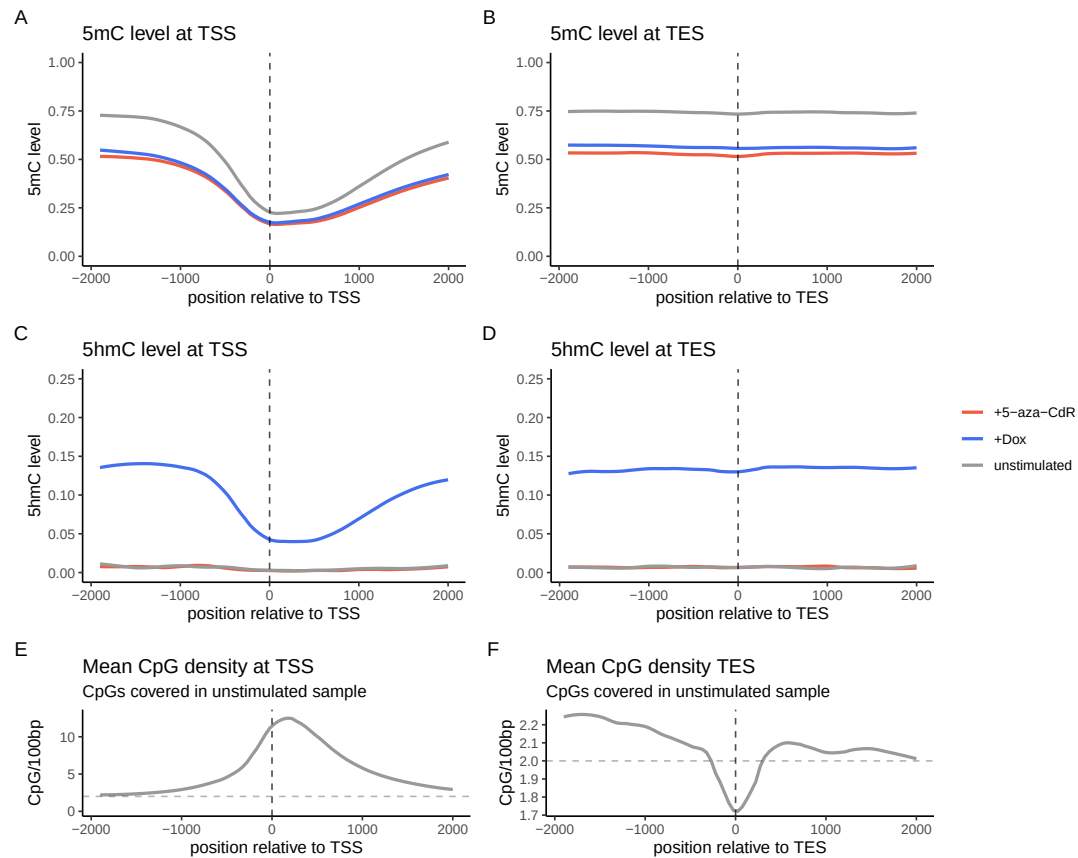


Figure 31: **DNA modifications in TSSs and TESs.** A-B show mean mC levels and C-D show mean hmC levels in TSS and TES, calculated across 100bp bins, summarised across genes. The CpG density, calculated as the number of covered CpGs per 100bp (both strands included) are shown in E and F.

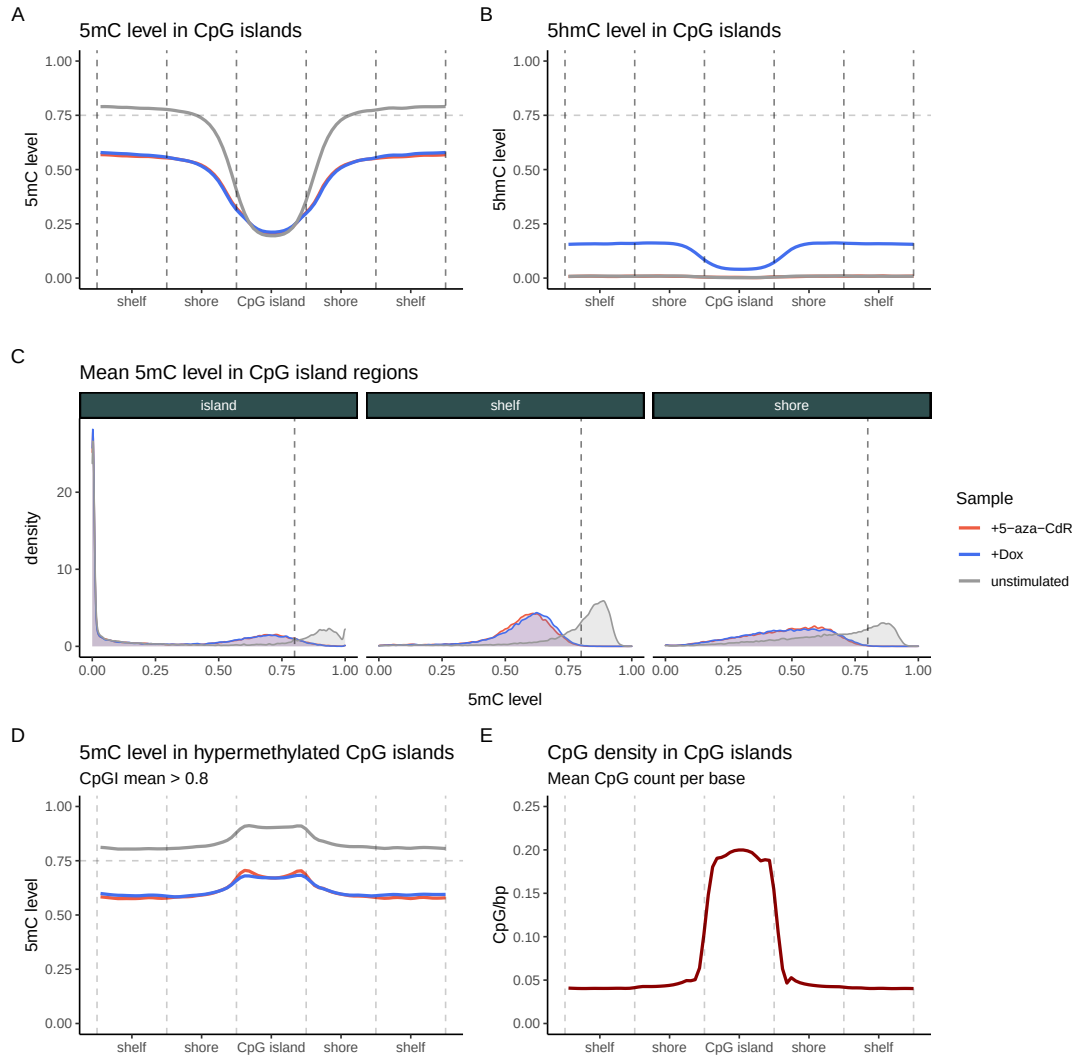


Figure 32: DNA modifications in CpGIs. A-B. Metagene plots summarise modification levels in the vicinity of CpGIs. mC and hmC levels calculated based on the TAPS and TAPS- β data are summarised into bins across CpGIs and adjacent shores and shelves (2k each). C. Density plots showing the mean mC level in CpGIs, shores, and shelves compared between conditions. D. Metagene plot showing the hypermethylated subset of CpGIs from A with mean mC level > 0.8 only. E. Average CpG density in CpGIs and flanking regions calculated as the mean CpG count per base.

and position-aware mC level per CpG.

We compared the mean mC levels in CpGIs between unstimulated and +5-aza-CdR, and found 78 examples of islands losing on average 0.5+. However, when viewing the regions in the genome browser, we found regions with mC diversity *within* the island, and inconsistent coverage between samples. In Figure 33, we see an example of a CpGI with significant mean mC loss in +5-aza-CdR, however it's clear that the mean signal in the CpGI is not a good measure of methylation since different CpGs are covered in each sample. Clearly coverage differences impact regional comparisons which do not require CpGs to be common between conditions. If the CpGs considered are found in both samples, this is not a problem. Therefore, rather than calculating mean mC per region first and subtracting to find mC loss, we calculate loss per position and use the mean of those values instead.

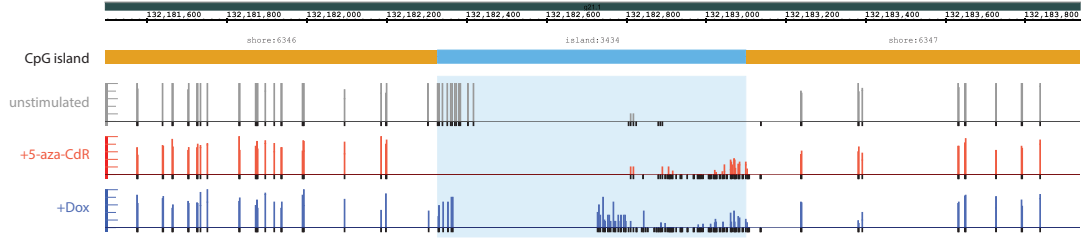


Figure 33: **CpGI methylation diversity.** The mC level per CpG as measured by TAPS- β in loci with 5+ coverage. Y-axis shows mC, 0 - 1, and black marks show covered positions in each sample (so that zero coverage and no mC can be distinguished). Only a fraction of the shores are shown and the island is highlighted.

To check whether this was indeed the reason for the lack of loss in Figure 32A, we examined individual CpG islands. Looking at a random subset of CpG islands, we can clearly note the prevalence of hypomethylation in the islands in the unstimulated sample and consistent reduction in the shelves and shores after +5-aza-CdR treatment (Figure 34). We then selected islands with a mean 5mC level of at least 0.8, and found a strikingly lower level of coverage within the islands (Figure 35). Indeed, when we restrict to fully covered CpGIs, we see the expected hypermethylation whereas summarising islands with low coverage results in an apparent prevalence of hypermethylation (Figures

36A and B).

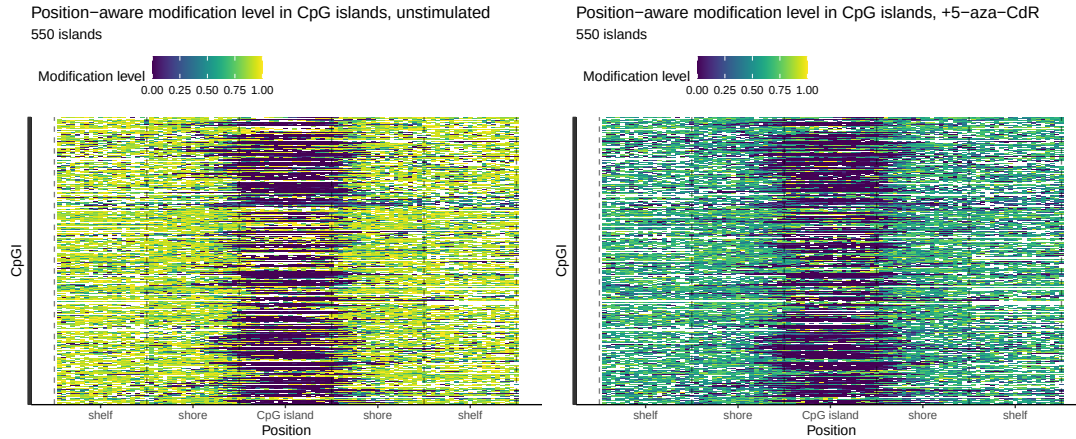


Figure 34: **Position-aware modification level in CpG islands.** A. Heatmap showing the mean 5mC level (color) in each bin (x-axis) across 550 randomly selected CpGIs (y-axis) for the unstimulated sample. Any missing data is shown in white. B. As A, but for +5-aza-CdR sample.

These sparsely covered islands can easily bias the results. We calculated the range of modification levels across all CpGIs and found that 13.7% show a range of mC levels over 0.7, in other words if we are unlucky enough to always pick the least methylated CpG in one sample but the most methylated in the other, we could be misclassifying a large proportion of CpGIs.

Using the per-CpG approach and filtering to have CpGIs with at least 75% positional coverage, we found that most of the CpGIs show no loss in mC (largely because they are unmethylated to begin with) (Figure 37A and B), which explains the lack of treatment effect in Figure 32A. We did find 234 CpGIs with average loss of 0.25+ in methylation level in at least one treatment and could thus isolate some examples of CpG islands with clear losses upon treatment (Figure 38).

To summarise, of the 25,837 CpGIs with at least one covered CpG, 18,947 CpGIs had 5+ covered CpGs in both samples and only 1 of those showed a 0.5+ mC gain after at least one treatment assuming that at least 75% of the island had some coverage. Therefore CpGIs are mostly unaffected by the demethylating treatments on account of being hypomethylated to start with, however this conclusion is dependent on the

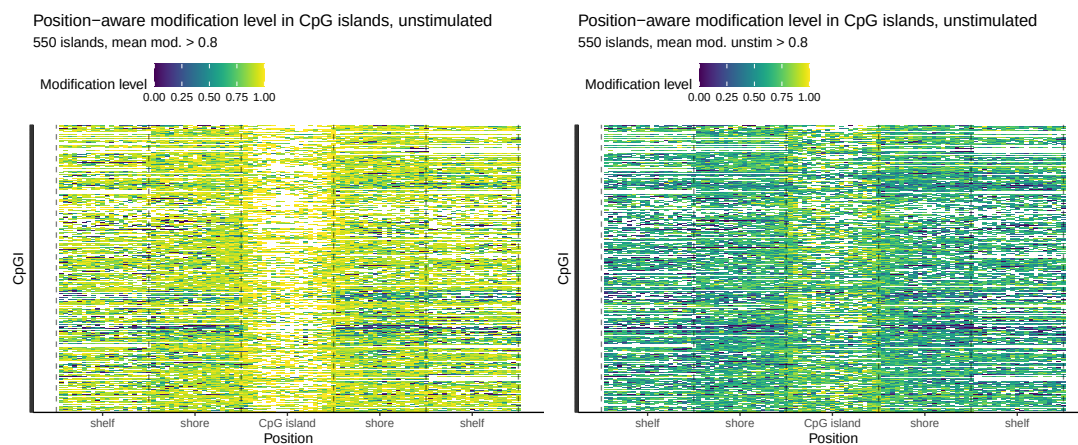


Figure 35: **Position-aware modification level in hypermethylated CpG islands.** As in Figure 34 but only CpGIs with mean 5mC level of at least 0.8 are included.

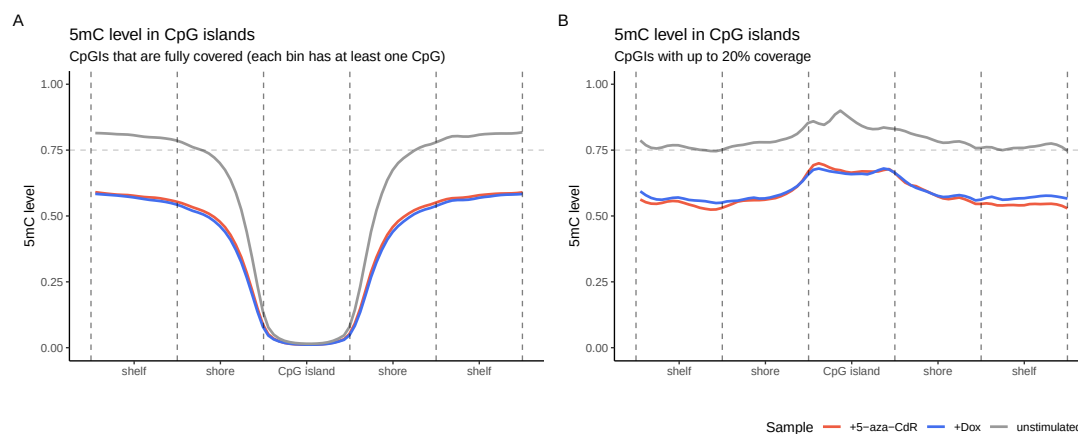


Figure 36: **Position-aware modification level in CpG islands split by coverage.** A. As in Figure 32A but only CpGIs with all bins containing at least one CpG included. B. As in A, but sparsely covered (1-2 bins of 20 only) CpGIs included.

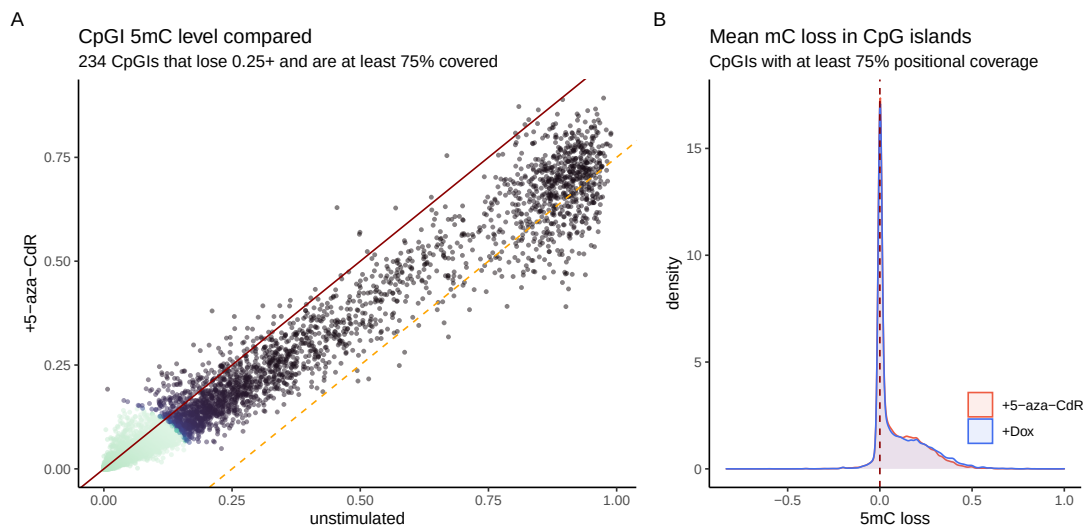


Figure 37: **CpGI mC loss.** A. Density scatter plot comparing the mean mC in unstimulated versus +5-aza-CdR treated. A line with slope 1 is shown in red, and another (in orange) with intercept -0.5 (such that the 78 points below represent CpGIs with mC level unstimulated $\geq 0.5 + \text{mC level +5-aza-CdR}$). B. Density plots comparing mean mC loss after +5-aza-CdR and +Dox treatment, calculated as mean of per-CpG mC loss for each CpGI.

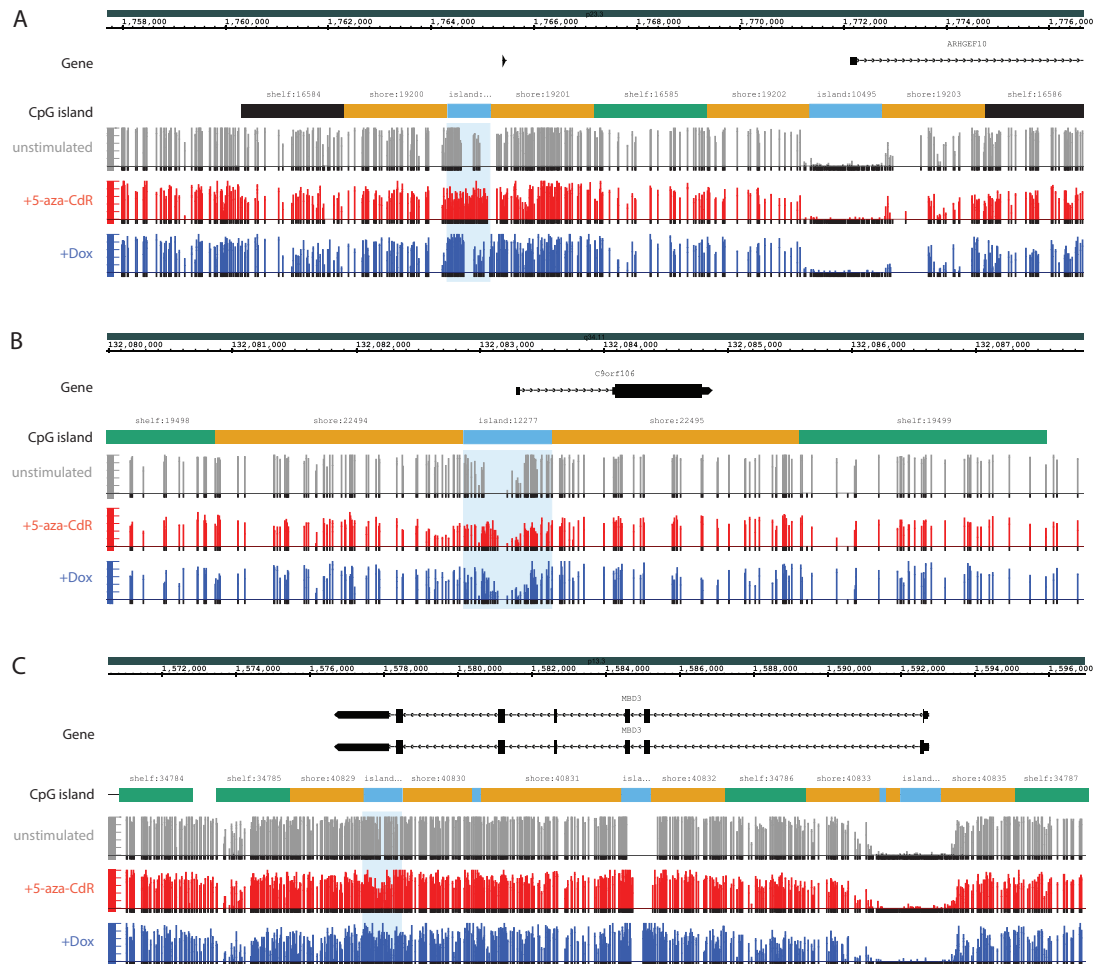


Figure 38: **CpGI mC loss in +5-aza-CdR**. 3 examples of CpG islands (highlighted) with highest mean mC loss after +5-aza treatment. See Figure 33 caption for further details.

summarising approach as we have seen.

3.3 Discussion

In this Chapter we wanted to confirm that the changes induced by +5-aza-CdR and Dox-induced TET1-CD are genome-wide, and largely unbiased. We designed a spike-in to measure hmC conversion in detail and have shown that interpreting conversion efficiency as modification level in circumstances of low coverage and/or incomplete conversion can lead to erroneous conclusions (and in our case chemically impossible³⁹ results). By modelling the effects of coverage and incomplete conversion we have shown that DNA modification values and changes calculated are consistent with the spike-in analyses, and made a case for using regional methylation estimates rather than focussing on individual CpGs to reduce the effects of variation⁴⁰.

Apart from increasing coverage and conversion efficiency, the system can be improved in multiple ways. One of the issues of using TAPS and TAPS- β to calculate hmC is the deconvolution of stochastic variation from true hmC abundance. Supposing the variances of the TAPS and TAPS- β samples at each CpG are σ_{mC} and σ_{mC+hmC} , then the variance of their difference is $\sigma_{mC} + \sigma_{mC+hmC}$. A method measuring hmC abundance directly would have variance σ_{hmC} , and as long as the conversion efficiency of hmC is comparable to TAPS (or TAPS- β) then this would be a significant improvement in accuracy. CAPS is a newly proposed related method to map hmC only which doesn't use TET [151]. Instead 5hmC is oxidised to 5fC by KRuO_4 before undergoing pyridine borane reduction. At the time of writing, this technique was still being optimised, but it would be very interesting to compare our estimates of hmC in this cell line with those obtained from using CAPS.

However, to avoid conversion inefficiency altogether, a different sequencing platform could be used. Oxford Nanopore Technologies (ONT) sequencing can measure mC and hmC simultaneously using ionic current⁴¹. Unfortunately, currently the accuracy of detection is quite low. Jain et al. [105] calculated an r-value of 0.895 using SignalAlign

³⁹As TAPS measures mC+hmC, and TAPS- β measures mC, one cannot have a negative values of TAPS - TAPS- β at 100% conversion since a negative hmC measurement is senseless.

⁴⁰As long as the variance is still a significant problem

⁴¹The website states that 5mC, 5hmC, 6mA, and BrdU in DNA have been measured.

[195] when comparing the ONT results with published bisulfite sequencing data and the hmC results are worse, with Rand et al. [195] reporting accuracy of 0.79 for hmC (with the same software). So, this approach could solve a lot of problems, but significant improvements to calling need to be made first.

Another aspect to consider is the mC loss we saw after TET1-CD upregulation. Similarly to above, measuring something directly rather than inferring effects is less affected by variation. We wanted to investigate the effects of hmC gain, however since the +Dox samples both lost mC and gained hmC, we saw significant overlap with the +5-aza-CdR results. To achieve hmC gain equivalent to mC loss, one can reduce the effect of passive demethylation by inducing cell-cycle arrest. Recall that DNMT1 copies mC, but not hmC, so the TET1-CD induced hmC's would propagate into the daughter cells as unmodified C (uC). Dr. Hiromi Tagoh used a double thymidine block to prevent replication in U2OS cells⁴² and found that the abundance of uC did indeed decrease (Figure 39). This system would however not work well with +5-aza-CdR, due to replication timing. During replication, the entire genome is accessible for DNMT1 which follows the replication fork copying mC onto the newly synthesised daughter strand. However some regions are considered early replicating (and others late), and were 5-aza-CdR added to cells with matched cell cycle then the late replicating regions would be more affected due to the progressive reduction of DNMT1, meaning we would have more demethylation at late replicating regions. This is not a problem when the population of cells is dividing at different times..

Some differences cannot be addressed easily, such as the slight bias for fully methylated mC in the +Dox treated samples. Recall that although there was a linear relationship between baseline mC and mC loss after +5-aza-CdR treatment, the +Dox treatment was less efficient at demethylating CpGs with intermediate baseline methylation. We proposed that one possible explanation for this may be differences in accessibility. Where +5-aza-CdR has equal access (on average, see discussion about late replication above) to each mC, TET1-CD is a protein that must access the methylated

⁴²These cells showed greater changes in expression after +5-aza-CdR and +Dox induction compared to HEK293 and therefore may be better candidates for further investigation. Discussed more in Chapter 5.

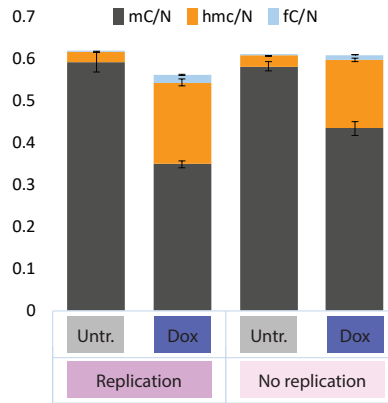


Figure 39: **MS results of cell cycle arrest.** Measurements by Dr. Paolo Spingardi.

CpGs and is therefore subject to constraints of chromatin accessibility. We ran a quick analysis of the relationship between CpG density and baseline methylation and found that CpGs with intermediate (0.3 - 0.7) modification level are more likely to be on their own. This observation can be supplemented with a combined analysis of DNA methylation and chromatin accessibility⁴³ to see just how much TET1-CD induced oxidation is dependent on chromosome state.

⁴³ATAC-Me is a method proposed by Barnett et al. [13] which can assess both simultaneously.

4 The effect of genome-wide DNA modification changes on occupancy of actively engaged RNA polymerase

4.1 Introduction

In this Chapter, our goal is to understand how DNA modifications affect RNA polymerase (RNAP) production rate through analysis of recently transcribed nascent RNA. Steady state mRNA levels are dependent on nascent RNA production rate, RNA processing, and degradation [111], thus, by measuring nascent RNA, we can isolate how perturbing DNA modifications affects transcription before downstream regulation can influence the RNA.

There is a curious gap in the field with few studies analysing nascent RNA synthesis and DNA modifications. Lorincz et al. [152] found that intragenic DNA modifications induce changes in the chromatin structure resulting in loss of elongation efficiency. Similarly, in Hetzel et al. [94]’s analysis of active RNAP occupancy in *Arabidopsis*, long (>2kb) genes exhibited TES-proximal CpG hypomethylation, which corresponded with strong pausing. However there is little data linking DNA modifications and active transcription, and in this chapter our target is to do just that.

With a working system of DNA modification perturbation in place, in this Chapter we first introduce our approach to measure occupancy of actively transcribing RNAP. Specifically we discuss the techniques used as well as the experimental design and quality control steps. We use an annotation agnostic analysis pipeline to check how DNA modifications affect initiation and elongation genome-wide, before moving on to a supervised analysis to detect gene candidates with changes in initiation, pausing, elongation, stalling, or termination.

4.2 Results

4.2.1 Experimental set-up

In lieu of separately measuring every RNAP per unit time, one can instead analyse the density of RNAP across the genome (Figure 40). This density will depend on the following (see Ehrensberger et al. [55]):

- How often RNAP initiates, i.e. initiation frequency. RNAP abundance on a gene is directly related to the number of polymerases that initiated transcribing that gene per unit time.
- The likelihood of RNAP remaining bound to the template throughout transcription. Since the elongating complex is relatively stable [60] we expect this likelihood to be similar for the majority of elongating RNAPs.
- Elongation rate also governs how the final density will look. This can vary between genes, and even within gene bodies. Without including a suitably small duration temporal component, differentiating between few slow-moving and many fast-moving RNAPs isn't feasible (Figure 40D). In our analyses, we analyse the data under the assumption that RNAP elongation speed (namely how quickly nucleotides are added) does not vary but the pausing stop-start dynamics can (similarly to Figure 40A). This remains a limitation of excluding a temporal component.

Precision run-on sequencing (PRO-Seq) captures recently transcribed, biotin-labelled nascent RNA through multiple rounds of biotin-selective pull-down, allowing for base-resolution mapping of active RNAP (see Section 1.3.3). Using this technique we measured the density of RNAP in a strand-specific manner.

The changes induced by the treatments had demethylating effects genome-wide (see Chapter 3) and mC is associated with repression, so we hypothesised some genome-wide changes in transcription. To differentiate between inconsistencies in library size and global changes in transcription, chicken fibroblasts were included in the library preparation as spike-ins, constituting roughly 10% of the cells (Figure 41).

Nuclei were isolated and incubated at 30degC for 30 minutes, during which RNAP transcribed in the presence of equal concentrations (0.05 mM) of ATP, GTP, biotin-11-CTP⁴⁴, and biotin-11-UTP⁴⁵. The biotin-labelled nucleotides would terminate transcription when incorporated, measuring active RNAP occupancy to the nearest C or U.

⁴⁴Jena Bioscience, NU-831-BIO

⁴⁵Jena Bioscience, NU-821-BIO

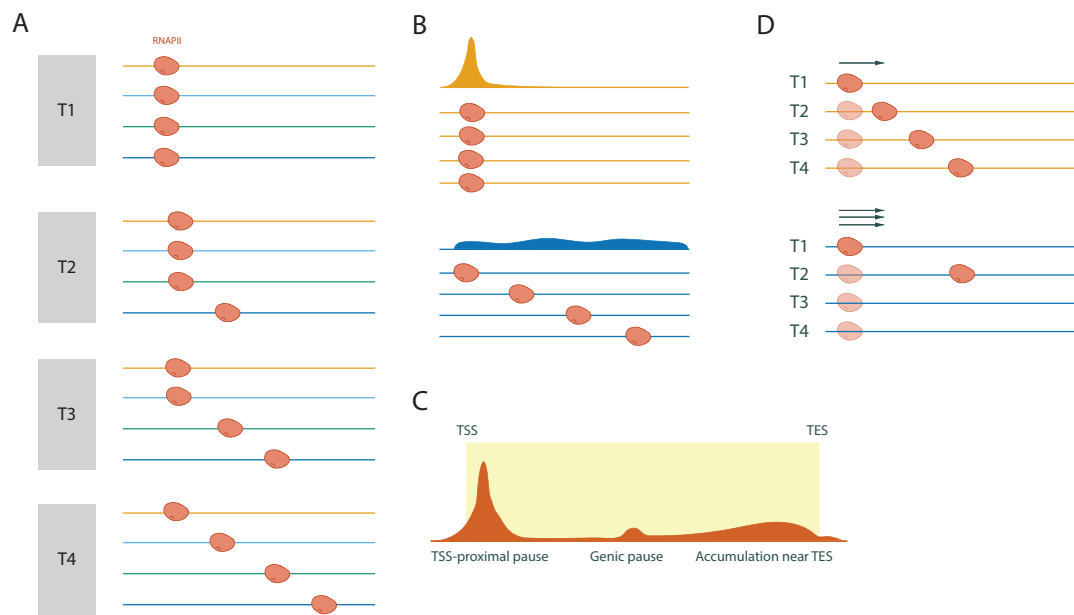


Figure 40: **Schematic to understand RNAP density.** A. Given four DNA regions, we note the position of RNAP at each of four consecutive time-points (T1-T4). Since the distance travelled (when transcribing rather than paused) is the same for each, in this case each has the same speed (but different pausing). B. Summarising over multiple molecules for each gene, the RNAP density can be used to infer what proportion of RNAP is paused genome-wide. D. The effect of different speeds can result in slow RNAP's dominating the signal.

TSS-proximal RNAP pausing is prevalent across vertebrate genomes, and would account for the majority of RNAP signal. To differentiate between TSS-proximal pausing and pausing elsewhere in gene bodies we used a complementary technique, PRO-Cap, to map initiation sites. In PRO-Cap, reads are additionally filtered to contain the 5' cap. The combination of these two techniques allowed us to analyse changes in initiation, pausing, and elongation.

To check reproducibility and robustness of results, each of the experiments was carried out in triplicate. This allowed us to measure experimental variability and add confidence to the results obtained.

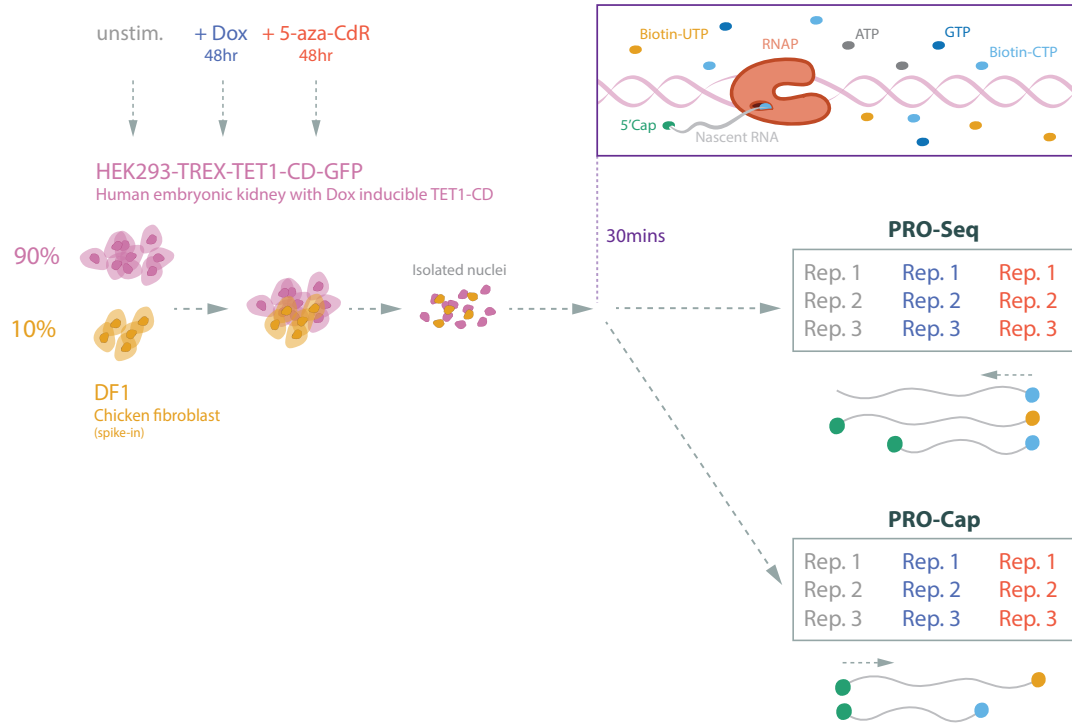


Figure 41: Experimental set-up. First, the HEK293 cells with inducible TET1-CD were treated with +5-aza-CdR or +Dox for 48 hours to modify mC and hmC levels (Section 3.2.1). The treated cells and one unstimulated control sample were then combined with chicken fibroblast cells (used as a spike-in) in a 9:1 ratio. The nuclei were isolated and RNAP was allowed to transcribe in the presence of ATP, GTP, biotin-CTP, and biotin-UTP for 30 minutes. For PRO-Seq, the RNA was fragmented before reads containing a biotin were isolated to map RNAP occupancy. In PRO-Cap, reads containing both the 5'-cap and biotin were selected to map the initiation sites.

4.2.2 Pre-processing and deduplication

The PRO-Seq, and more so PRO-Cap, libraries are subject to moderate duplication levels since some (start, end) coordinate pairs are significantly more likely to occur compared to others. For example, assuming a well-defined TSS at position X and pausing locus at Y , we expect to see multiple $[X, Y]$ reads which show that RNAP is paused at Y after starting at X in multiple cells. If we were to deduplicate these reads directly, we would underestimate RNAP abundance at well-defined TSSs and pausing sites. Therefore, to ensure that as many distinct RNAP's as possible could be mapped (even if they occupied the same locus) unique molecular identifiers (UMIs) were included during library preparation to allow for downstream deduplication.

UMIs are short, unique sequences which can be incorporated into the start and/or end of a DNA or RNA fragment of interest to aid deduplication after PCR by matching each read with a molecule-specific tag. In our experiments, an 8bp 5' UMI and 4bp 3' UMI were added to the PRO-Seq reads, and 4bp 5' UMI and 4bp 3' UMI were added to the PRO-Cap to assist deduplication (Figure 42)⁴⁶.

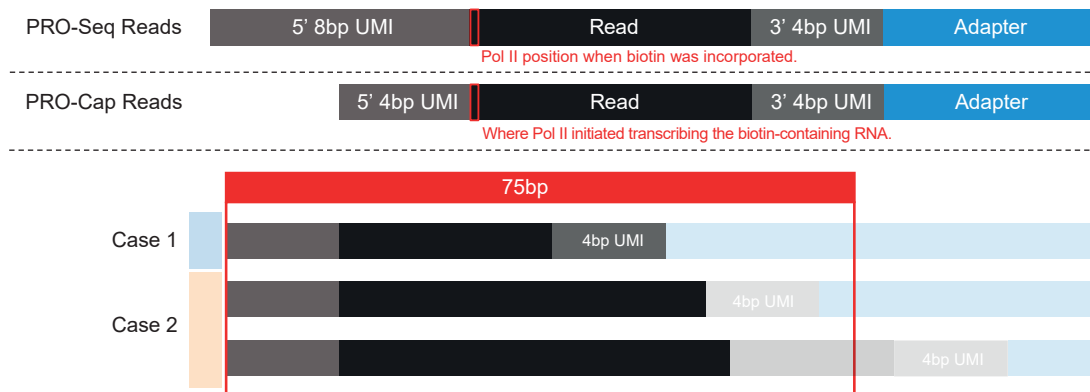


Figure 42: **Unique molecular identifiers (UMIs) in PRO-Cap and PRO-Seq.** PRO-Seq reads were prepended by an 8bp UMI, and appended by another 4bp UMI; the 5' 8bp UMI was then used to deduplicate the reads. PRO-Cap reads were prepended by a 4bp UMI, and appended by another 4bp UMI; for reads with 3' adapter content exceeding 5bp, the 3' 4bp UMI was appended to the 5' 4bp UMI, creating an 8bp UMI.

⁴⁶Using a 4bp 5' UMI in the PRO-Cap data was an oversight and in future experiments, an 8bp 5' UMI would make the downstream analysis much easier.

The base distributions clearly show where the UMIs end in both experiments (Figure 43). A clear overrepresented adenosine appeared in the PRO-Cap base distribution following the 4bp UMI, coinciding with the initiator (Inr) element associated with transcription ($C_{-1}A_0$), [125], and the high GC content near the TSS was consistent with what would be expected in vertebrates. The PRO-Seq base distribution downstream of the 5' 8bp UMI suggested that biotin-CTPs were most likely to halt elongation (since the complementary guanine is overrepresented). Even though both biotin-CTP and biotin-UTP were used, biotin-UTP was either less efficient at causing RNAP to terminate or less likely to be incorporated by RNAP⁴⁷. This means that RNAP occupancy at cytosines will be over-represented and the data should be interpreted as the RNAP position calculated to the closest downstream cytosine.

The mean duplication in the raw data was 5.24% and 2.16% in PRO-Cap and PRO-Seq, respectively. Higher duplication in PRO-Cap is expected as the sequenced library is a subset of the library available for PRO-Seq and the UMI length is shorter. To deduplicate the PRO-Cap data, we modified the pipeline since the 4bp 5' UMI can only differentiate between 256 groups (assuming equal distribution). We noted substantial adapter content in PRO-Cap samples in particular suggesting that most of the RNAPs paused within 75bp of the TSS before elongating (Figure 43)⁴⁸. Since most PRO-Cap reads in each replicate contained adapter content exceeding 15bp (Figure 43), we increased the length of the PRO-Cap UMI using the 4bp 3' UMI.

We found the length of each PRO-Cap read after adapter trimming and split reads into:

- *short*: those with adapter content exceeding 5bp, i.e. reads shorter than 69bp
- *long*: everything else

The *short* reads were selected for UMI concatenation (Case 1 in Figure 42). We appended the 4bp directly preceding the adapter to the original 4bp 5' UMI, then

⁴⁷The frequency of A (complement to U) in PRO-Seq is quite low and didn't match T, suggesting that RNAP may have struggled to incorporate biotin-UTP.

⁴⁸This is consistent with previous findings of RNAP pausing, and provides some justification for using single-end sequencing as most reads would substantially overlap if using paired-end.

trimmed the read by 4bp accordingly. Thus we extended the length of the UMI to 8bp (differentiating between 65,536 possible groups) before deduplication.

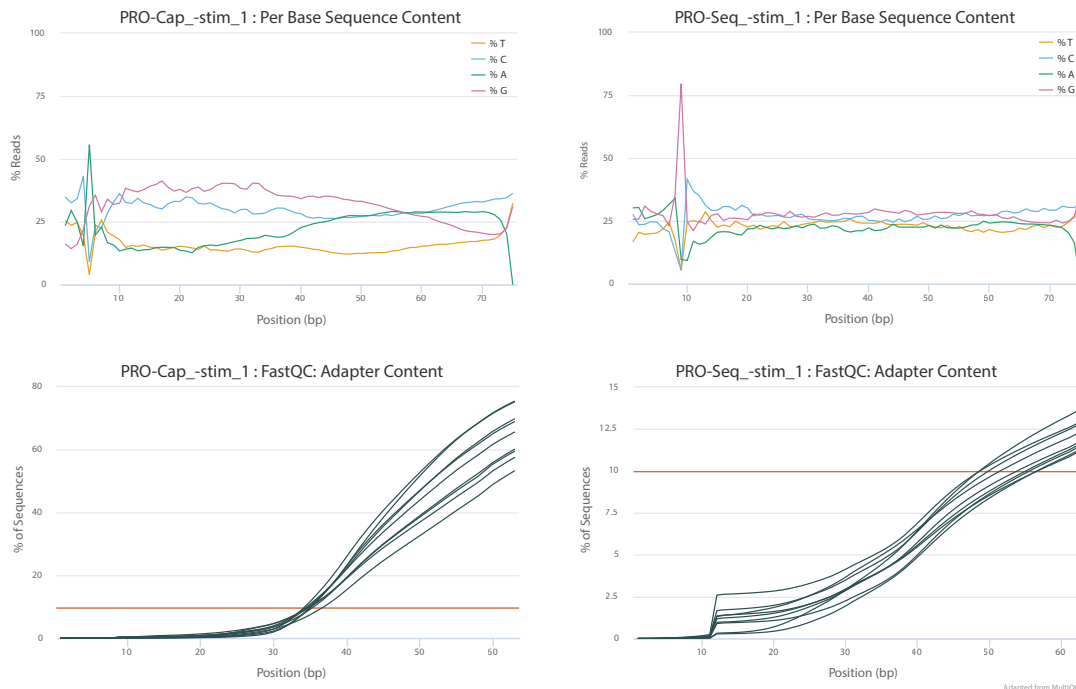


Figure 43: **Raw PRO-Cap and PRO-Seq base distribution and adapter content.** The per base sequence content of a representative PRO-Cap and PRO-Seq sample calculated by FastQC [5]. The bottom panels show the cumulative percentage of sequences with adapters starting at various positions. Original plots generated by multiqc [62], were combined and adapted to be colourblind-friendly.

4.2.3 Sample normalisation and spike-ins

We incorporated chicken nuclei as a spike-in into each of our run-on experiments to aid normalisation in the event of global changes in RNAP occupancy post-treatment. How the information from spike-ins is incorporated into the analysis has varied in the literature:

- Booth et al. [24] used a normalisation factor of spike-in mapped reads per 10,000, similarly Ball et al. [10] used a library size correction factor (sample library size divided by mean library size) and spike-in correction factor (ratio of spike-in DNA) to normalise the read count data.

- Dukler et al. [52] did not consider library size directly, and used the ratio between sample spike-in reads and the median spike-in reads across samples.
- In lieu of spike-ins, Zhang et al. [270] considered reads from the mitochondrial chromosome as a control and downsampled their data accordingly.

We defined the normalisation factor as galgal5 (chicken) signal as a proportion of hg19 signal. We saw that spike-in proportion varied between batches, but not strongly *between* conditions, and normalisation had potential to inflate the counts in batch 3 in particular (Figure 44). Therefore, similarly to Elrod et al. [60], we decided that since spike-in recovery was similar *between* conditions, the data should be normalised using DESeq2 size factors instead.

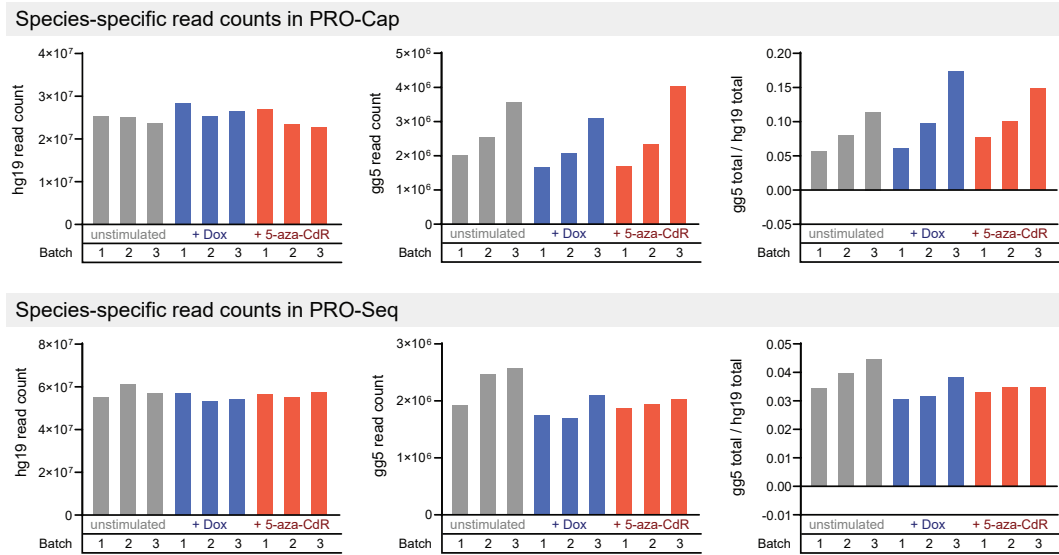


Figure 44: **Spike-in normalisation.** The number of PRO-Cap (top) and PRO-Seq (bottom) reads mapping to hg19 and the spike-in, galgal5, per sample. Normalisation factors are calculated as a ratio between galgal5 read count and hg19 read count.

4.2.4 Overview of initiation and occupancy of active RNAP in HEK293 cells

To check accuracy of proposed initiation and elongation regions relative to available annotation (and thus whether the experiments worked well), we first looked at the data

in the genome browser (Figure 45). Indeed, PRO-Cap appeared to coincide with annotated TSSs in most cases, with PRO-Seq showing evidence of TSS-proximal pausing followed by elongation in many genes. Some genes were more active than others (compare MAGI3 and LRIG2), while SYCP1 showed no active transcription at all. We often saw peaks mirrored on both strands, indicating prevalence of TSS-proximal antisense and divergent transcription.

The data looked promising, so we calculated metagene profiles⁴⁹ for comparison with published data (Figures 45B and C). The PRO-Cap signal shows a strong sense peak at the TSS (with some evidence of initiation downstream) and substantial antisense and divergent promoter-proximal signal as described by Core et al. [39] and Nojima et al. [176]. Considering GENCODE annotations were used to define gene boundaries in this case, the enrichment near the TSS suggests that most major data-driven TSSs are close to GENCODE-annotated TSSs for this cell line.

The RNAP occupancy map mirrors the PRO-Cap at the TSS, but with larger signal downstream of the TSS, showing evidence of widespread pausing. The majority of engaged RNAPs are found near the TSS and a small proportion are undergoing active elongation across the gene body, consistent with the data presented by Steurer et al. [228].

We noted mostly sense signal beyond the TES, consistent with previous findings [176]. However, the dip just upstream of the TES seemed suspicious, so we calculated the base distribution around the TESs and found that there is an under-representation of C's ~23bp upstream of the TES (Supplementary Figure S1) which explained with the drop in signal⁵⁰.

Based on this annotation-driven analysis we were convinced that the PRO-Cap data was credible and could indeed map initiation sites accurately. The PRO-Seq signal also behaved as expected relative to the PRO-Cap and annotation.

⁴⁹Non-overlapping genes only.

⁵⁰Recall PRO-Seq is measuring RNAP position to the nearest cytosine.

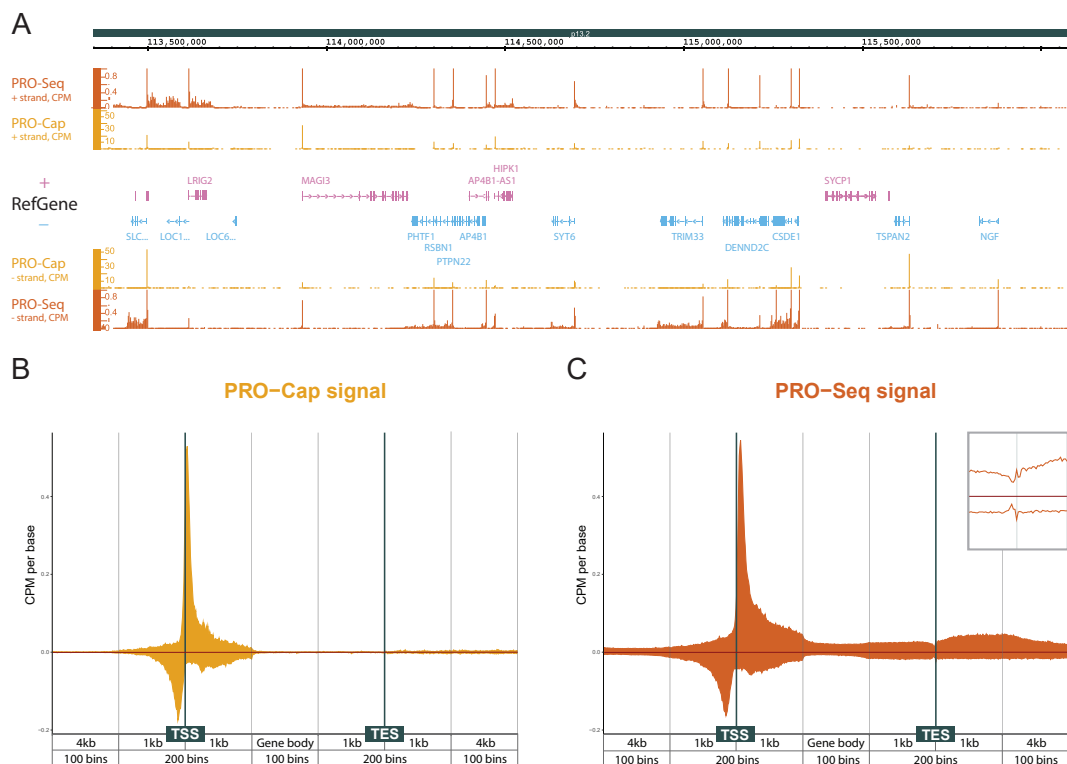


Figure 45: **PRO-Cap and PRO-Seq profiles in unstimulated HEK293 cells.** A. Genome browser view of typical PRO-Cap and PRO-Seq signal relative to gene annotation (RefSeq). The reads are shown per strand, summarised over the three unstimulated samples for each experiment. Genes are coloured according to sense strand, pink for plus and blue for minus. B-C. PRO-Cap and PRO-Seq density profiles summarised over non-overlapping genes for each condition. The sense signal is shown as positive density, and antisense is shown as negative density. We normalised using counts per million (CPM) for each replicate before calculating mean depth per position. In addition, the PRO-Seq panel contains a zoomed in view of the discussed dip in RNAP density at the transcription end site (TES).

4.2.5 Are RNAP initiation and occupancy metagene profiles affected by DNA modification changes?

We compared the PRO-Cap and PRO-Seq metagene profiles between conditions to ascertain if there were any clear differences after DNA modification changes (Figures 46A and B). The difference between conditions was minimal. We did note slightly higher TSS-proximal signal in the control, but it was difficult to interpret without a more detailed analysis. Given the data we have, this annotation-based approach can be improved significantly by defining data-driven TSSs (since the annotation usually contains the longest transcript for each gene).

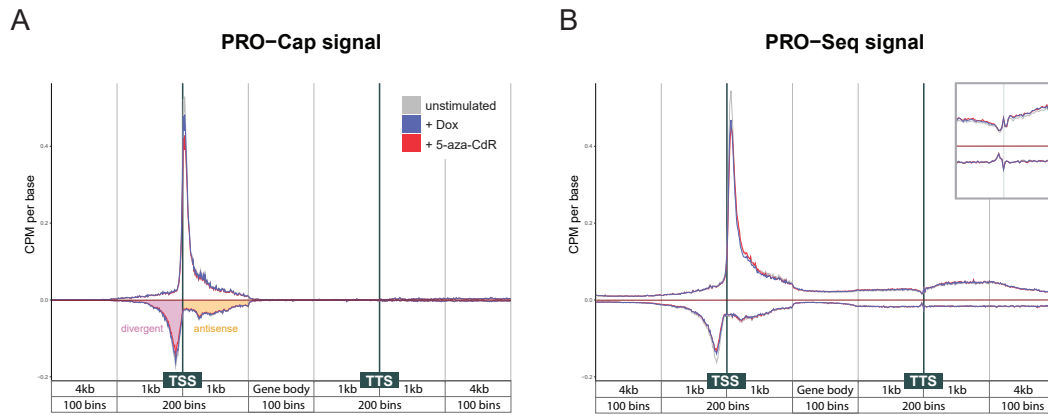


Figure 46: **Metagene PRO-Cap and PRO-Seq profiles.** PRO-Cap and PRO-Seq density profiles summarised over non-overlapping genes for each condition, focusing on the TSS and TES. The sense signal is shown as positive density, and antisense is shown as negative density. In order to compare the conditions directly, we normalised using counts per million (CPM) for each replicate before calculating mean depth per position. The PRO-Seq panel contains a

4.2.6 Annotation Agnostic Analysis

Multiple studies have linked changes in DNA methylation to spurious initiation. In this section we wanted to investigate whether the samples showed evidence of changes in RNAP initiation upon treatment with +5-aza-CdR or TET1-CD upregulation.

We did not expect to see strong genome-wide increases or decreases in PRO-Cap or PRO-Seq signal based on the spike-in analysis, however redistribution of reads was a

possibility, and we decided to first explore this in an unsupervised manner, i.e. independent of genome annotation, through:

- A binned analysis comparing the read distributions in down-sampled replicates is a crude way to analyse if the initiation sites were distributed differently across the genome.
- A peak analysis comparing the number of initiation regions in each condition would allow us to calculate peak stability more directly by examining variation between replicates, as well as the distribution of initiation peaks.

4.2.7 Binned initiation

Starting with the most simple approach, we binned the genome into 100bp non-overlapping bins and quantified the frequency of read counts per bin for each downsampled replicate. The read distribution in the samples appears different between conditions⁵¹ (Figure 47A). By comparing the frequencies of 100bp bins with 1, 2, 3, etc. reads, we note that the +5-aza-CdR and +Dox samples have more low-abundance bins compared to unstimulated, which suggests a wider distribution of signal across the genome. Indeed, if we compare the bin count frequencies we can see a systematic shift with +5-aza-CdR and +Dox dominating low-count bins in the downsampled data (Figure 47B). These observations led us to hypothesise that the treated samples may have more evidence of spurious initiation.

4.2.8 Data-driven initiation

4.2.8.1 HMM model

To evaluate whether magnitude (size of peak) or frequency (count) of initiation is affected by changes in mC and hmC, we wanted to identify data-driven TSSs based on the PRO-Cap data and compare these initiation sites between conditions. Some unsupervised methods used for identifying initiation regions involve binning the genome and assuming that the counts follow a Poisson distribution [230, 91], but since Poisson

⁵¹But the differences were not statistically significant.

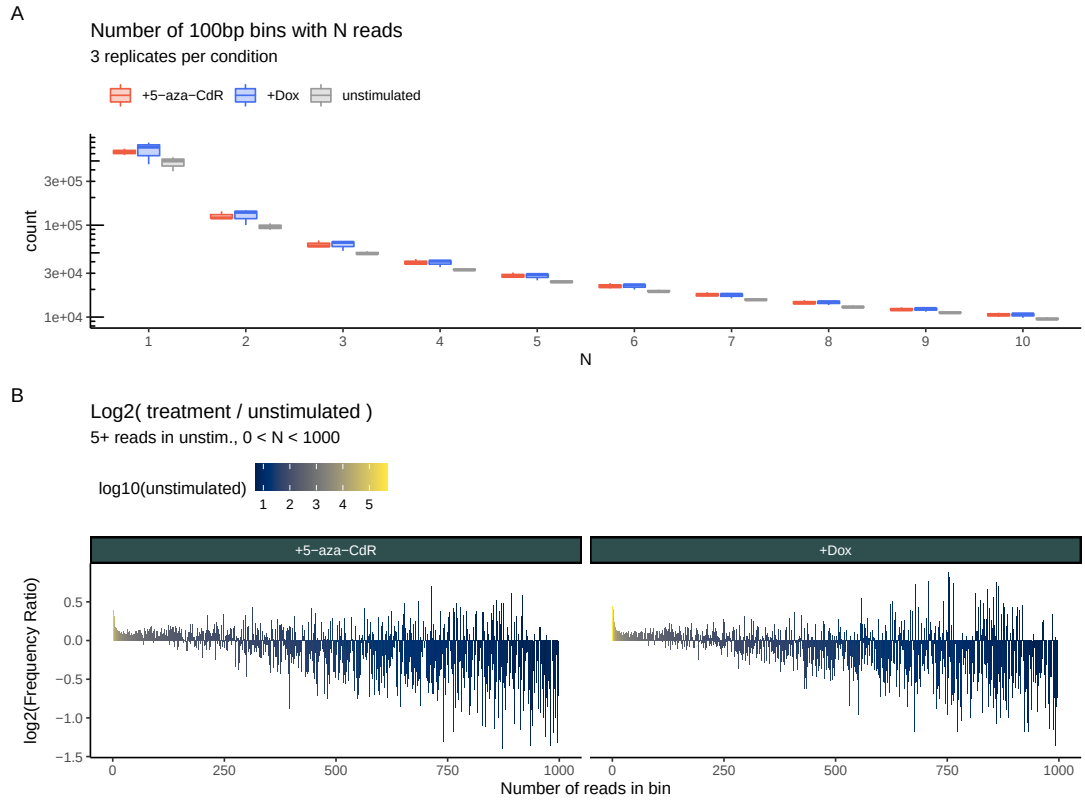


Figure 47: **Read distribution analysis.** A. The number of bins with N reads, grouped per condition (this shows a subset of the N's in B). B. Ratio of mean bin count frequencies. E.g. for $N = 5$, we compare the mean number of bins in unstimulated with 5 reads and the mean in +5-aza-CdR.

assumes independence between neighbouring bins we decided to look for alternative approaches.

Hidden Markov Models (HMMs) allow for greater tuning and incorporation of neighbouring data. Broadly, a HMM classifier attempts to assign a “hidden state”⁵² based on observed states and pre-defined knowledge about:

- the initial state probabilities, e.g. given a 10bp bin at the start of a chromosome, what is the probability that it is background
- the transition matrix, containing the probabilities of moving between states, e.g. $P(\text{background} \rightarrow \text{peak})$, $P(\text{background} \rightarrow \text{background})$, etc.
- the emission matrix⁵³, containing the probabilities of seeing each hidden state given the observed state, e.g. if we consider a bin as “high” whenever the signal in that bin exceeds 10 reads then $P(\text{background} \rightarrow \text{high})$ is specified in the emission matrix (and is likely to be low)

HMMs have been used to call initiation regions in multiple studies [40, 171, 228]. We implemented a modified version of the model used by Core et al. [40], which was designed to differentiate between narrow and broad peaks, making it suitable for calling both (see transition probabilities in Figure 48).

4.2.8.1.1 HMM design

One of the advantages of PRO-Cap is the generation of bp-resolution data, however based on the distribution of reads in the genome browser we noted that the initiation sites usually clustered together in groups but are rarely specific to a single base, consistent with other studies describing clusters of TSSs [179, 82]. Therefore, to reduce the computational burden of calling peaks in the PRO-Cap data without losing much granularity, we summarised the counts in non-overlapping 10bp regions.

Peaks were called using a modified 3-state Hidden Markov Model (HMM) (see Section 2.4.4.1). The algorithm classifies each 10bp bin into background (B), broad peak

⁵²Something that we cannot measure directly, for example whether a 10bp bin belongs to an initiation peak.

⁵³There can be multiple observed states.

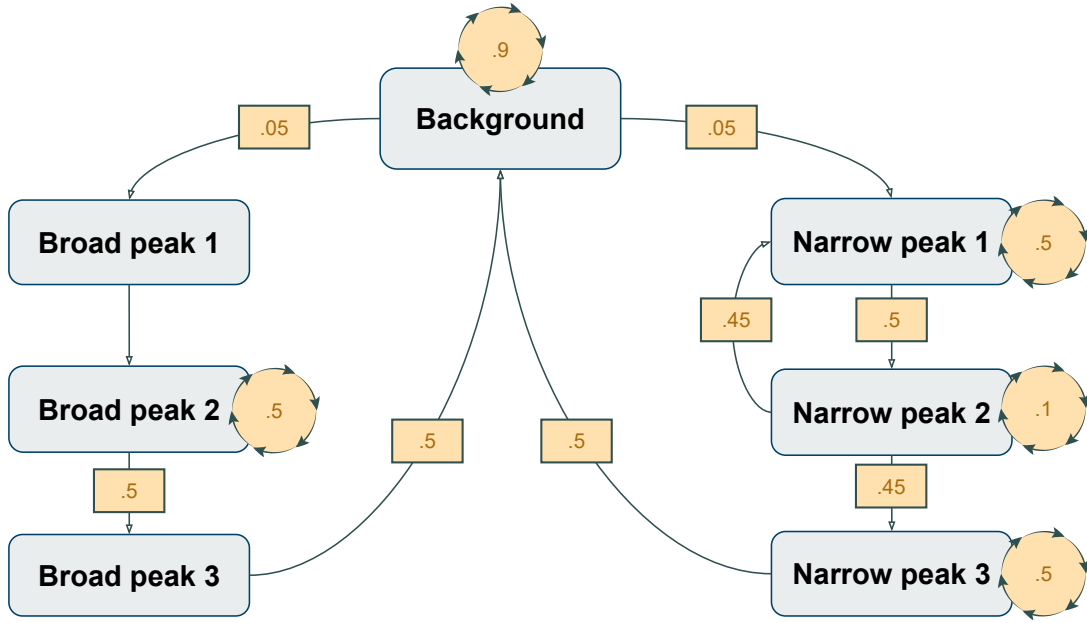


Figure 48: **HMM transition probabilities.** The model shows the probabilities of staying in the same state (circles) or moving to a new state (rectangles). The model attempts to classify regions into background, narrow peaks, and broad peaks (with the latter split into components). See additional details in Section 2.4.4.1.

(BP), or narrow peak (NP), based on two observed values: the signal and the peaked-ness. Let us annotate the non-overlapping 10bp bins in a region within a chromosome as b_i , where b_i and b_{i+1} are adjacent, and the read count in b_i is n_i . The signal for b_i is then classified as follows:

- empty if $n_i = 0$
- depleted if $0 < n_i < S$
- enriched if $S \leq n_i$

where S is a user-defined count threshold (we used 10). To calculate the peaked-ness of b_i , first the background signal is estimated as the mean signal in any depleted bins within 100bp of b_i (if two or more are found)

$$\begin{aligned}
bg_i &= \text{mean}(B_i), \text{ if } |B| \geq 2 \\
B_i &= \{b_j | j \in [i - 11, \dots, i - 1, i + 1, i + 11], 0 < n_j < S\}
\end{aligned} \tag{1}$$

then, the bin b_i is peaked if $fc \times bg_i < n_i$, where fc is a user-defined fold-change requirement. The signal and peaked-ness are the observed variables which are used to predict the hidden state.

4.2.8.2 Trimming

The HMM peaks often included regions of low signal near the boundaries highly enriched peaks, so we decided to trim them. In their analysis, Strbenac et al. [230] trimmed any outermost positions with zero counts before merging any peaks within 30bp. Core et al. [40] similarly trimmed any “depleted” intervals, however we wanted to trim using a prior based on signal in the surrounding area (rather than a single threshold).

Trimming requires two parameters, required proportion of the nearest maximum enrichment 10bp bin m_n and distance to search for that bin, d . For 10bp bin b_i in peak p_k , we first merge any peaks within dbp of p_k , call this search area P_k . We then define $m_k = \max\{n_j | b_j \in P_k\}$ as the count in the most enriched bin. Finally, we select the most enriched bins, $B_E = \{b_i | n_i \geq m_n \times m_k, b_i \in p_k\}$. To reduce some granularity, peaks within 50bp are merged. This final trimmed set will be referred to as *primary* peaks.

4.2.8.3 Stringency and variation analysis

We decided to vary the sensitivity of the caller by testing multiple fc and m_n (mxn) thresholds to evaluate the influence of these thresholds and assess replicate variability.

Independent of changes to the HMM and trimming parameters, we consistently saw slightly more peaks in the +5-aza-CdR and +Dox samples but the differences were not statistically significant (Figure 49A). Differences are most prominent at lower stringency levels, suggesting that spurious “low-count” peaks are more prevalent in the treated samples (consistent with binned analysis). We checked what proportion of peaks are

threshold-specific based on one of the replicates and, of the peaks called at the lowest stringency (fc2, mxn0.2), $\sim 58\%$ are accounted for by the highest stringency group (fc20, mxn0.7) (Figure 49B)⁵⁴. We identified low stringency peaks that were missing from the high stringency set and found that most (89.6%) were found within 1kb of high stringency peaks, with the remainder falling disproportionately into intergenic regions (Figures 49C and D). After comparing different threshold options in the genome browser we chose to use the fc10, mxn0.3 peak set in further analyses, however it's clear from the overlaps (Figure 49B) that small changes to these levels would not affect the final peak set (and conclusions) much.

We measured peak stability⁵⁵ by comparing calculated peak positions between replicates. We expected to see less variation between the unstimulated replicates compared to treated, since the randomly distributed hmC gain and mC loss had the potential to induce small changes in initiation independently at each CpG. We used the proportion of pair-wise non-overlap between replicates as a dissimilarity metric and found that the intra-condition dissimilarity is lower in the unstimulated samples, +Dox and +5-aza-CdR samples having a higher proportion of peaks than unstimulated (Figures 50A and B). To find whether these non-overlapping peaks were reproducible, we analysed the condition- and sample-specific peaks from each replicate. Although there were indeed more unique peaks in the +5-aza-CdR and +Dox replicates on average, the variation between replicates was too great to definitively conclude that the demethylating treatments result in less stable peaks (Figure 50C).

4.2.8.4 Pausing

Pausing near the TSS is a genome-wide phenomenon, and due to the substantial signal in pausing regions we wanted to check the consistency of pausing location, namely the downstream distance from the TSS⁵⁶. Since the majority of peaks overlapped between samples we chose to combine them and use the union of all called peaks to ensure that

⁵⁴Not all the highest stringency peaks are found in the lowest stringency group (only 91%) as a consequence of stronger thresholds reducing peak size.

⁵⁵We consider a peak “stable” if it occurs in multiple replicates (is reproducible).

⁵⁶By not removing the full pausing region from elongation data we may bias the elongation counts.

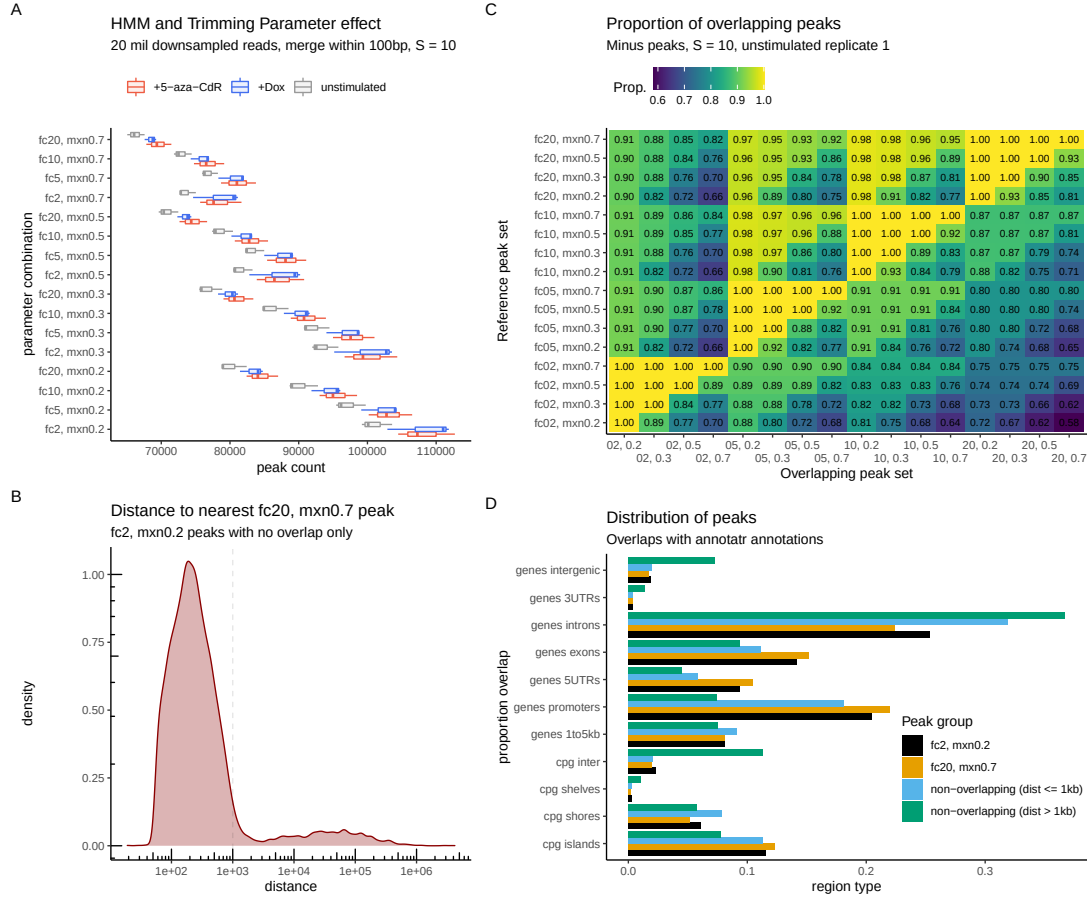


Figure 49: **Peak annotation parameter effect.** A. How different combinations of thresholds used affect the final number of called peaks. B. Proportion reference peaks that overlap with the overlapping peak set. C. Density of distances of low stringency (fc2, mxn0.2) peaks to high stringency (fc20, mxn0.7) peaks, for the 42% of low stringency peaks that are missing from high stringency group. D. The proportion of peaks overlapping annotatr [32] hg19 annotations.

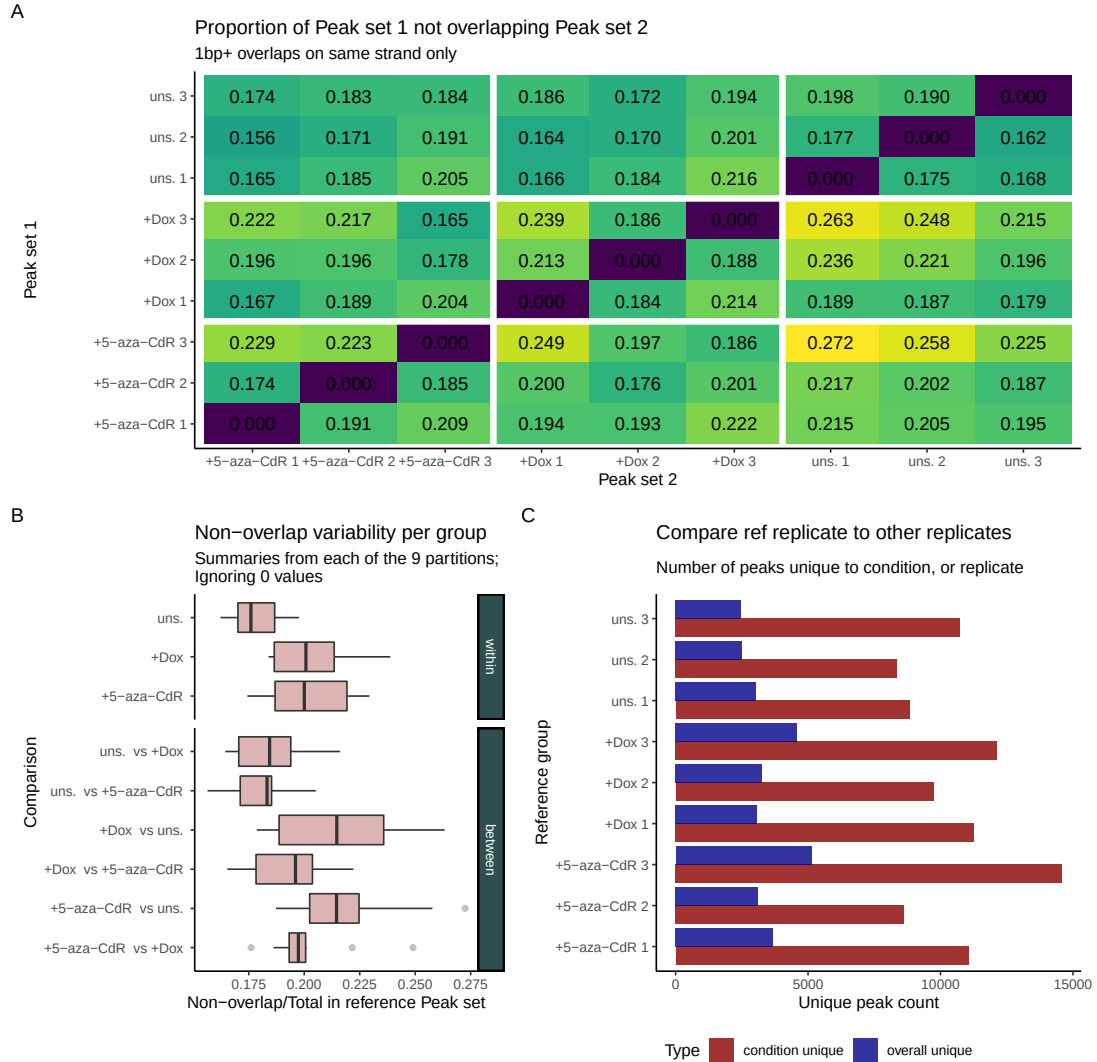


Figure 50: PRO-Cap Replicate variability. A. Heatmap showing proportion of peaks in Peak set 1 not overlapping a peak in Peak set 2. In all cases peaks are called on the downsampled PRO-Cap data with $S=10$, $FC=10$, $mxn=0.3$. B. Boxplots showing the distribution of values in the heatmap, with self-comparison overlap proportions excluded. C. Bar plots showing the number of peaks in each replicate that are unique to a condition (red): the peak may be found in other replicates from that condition only; or replicate (blue): the peak is only found in that replicate and has not been reproduced in any of the other samples.

the regions of interest were consistent between replicates⁵⁷. Merging resulted in 93,620 peaks across all replicates, call these the annotation agnostic TSSs (aaTSSs).

The pausing region (PR) of interest was defined as the 300bp downstream of the aaTSSs⁵⁸. We calculated the PRO-Seq read density in each PR across non-overlapping 10bp bins, and saw a clear enrichment of signal within the first 60bp downstream of the PRO-Cap peak as expected (Figure 51A). To evaluate profile consistency we split data according to total PRO-Seq count in the 300bp window of interest (Figure 51B), and then normalised the region counts to the proportion of PRO-Seq reads falling in each window downstream of the PRO-Cap defined TSS. The apex of pausing did indeed vary within each group (Figures 51C-E), however substantial pausing signal more than 70bp downstream of the PRO-Cap peak was rare.

We compared the pausing profiles between the conditions to see whether mC and/or hmC deregulation influenced pausing distance. Two of the three replicates suggested that +5-aza-CdR and +Dox treatments resulted in more distributed pausing on average, but this pattern was not as clear in batch 1 (Figure 52A). The most prominent difference was seen in batch 3, but when looking at random subsets of the profiles we did not see any clear differences which could have been masked by averaging (Figures 52B-D). Due to the variability between batches we cannot conclude whether the treatments indeed caused a global change in pausing distance.

4.2.8.5 Run-through

Not all aaTSSs are necessarily active, and we saw in the last section that many 300bp downstream regions contain 1-10 reads (Figure 51B). Here we would like to ascertain how often run-through signal makes up a large proportion of the elongation signal, where run-through signal is defined as elongation signal which started from an aaTSS upstream of the closest aaTSS. In Figure 53 we see an example of a TSS whose downstream region is likely to be affected by run-through signal from an upstream TSS, if this were a common phenomenon then the elongation signal would need to be adjusted based on

⁵⁷We considered a “present in two replicates” rule whereby any peak found in at least two replicates of a condition would be included. However the downstream analysis filtered out any batch-specific peaks due to inflated standard errors regardless.

⁵⁸We ensure that the initiation peaks are at least 500bp apart.

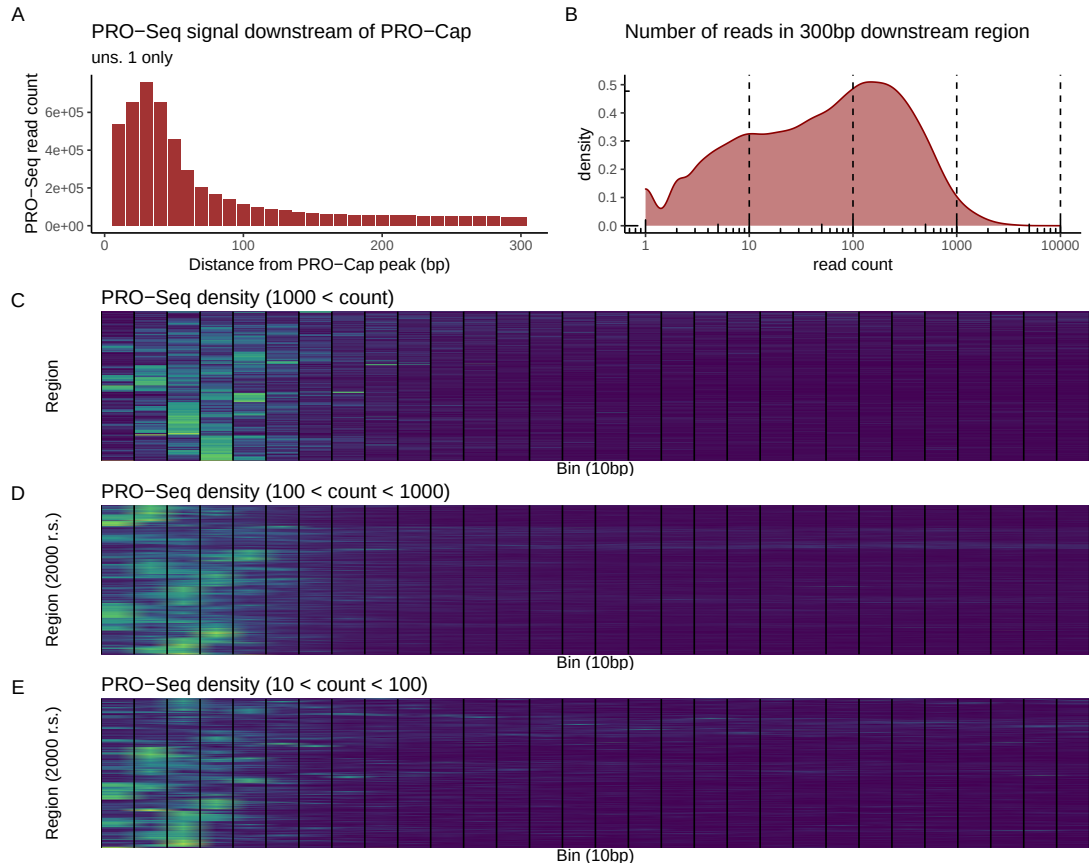


Figure 51: **PRO-Seq read distribution downstream of PRO-Cap peaks.** A. Bar plot showing the total read count in each 10bp bin downstream of a PRO-Cap peak in unstimulated sample 1 (largely representative of the other samples). B. Density of total read count in the 300bp windows of interest. C-E. Proportion of reads falling into each 10bp bin for different read count bounds, regions clustered together using complete linkage and euclidean distance in each case. For the larger sets a random sample (r.s.) of 2000 regions is used. Lowest values are in blue, and highest proportions are shown in yellow.

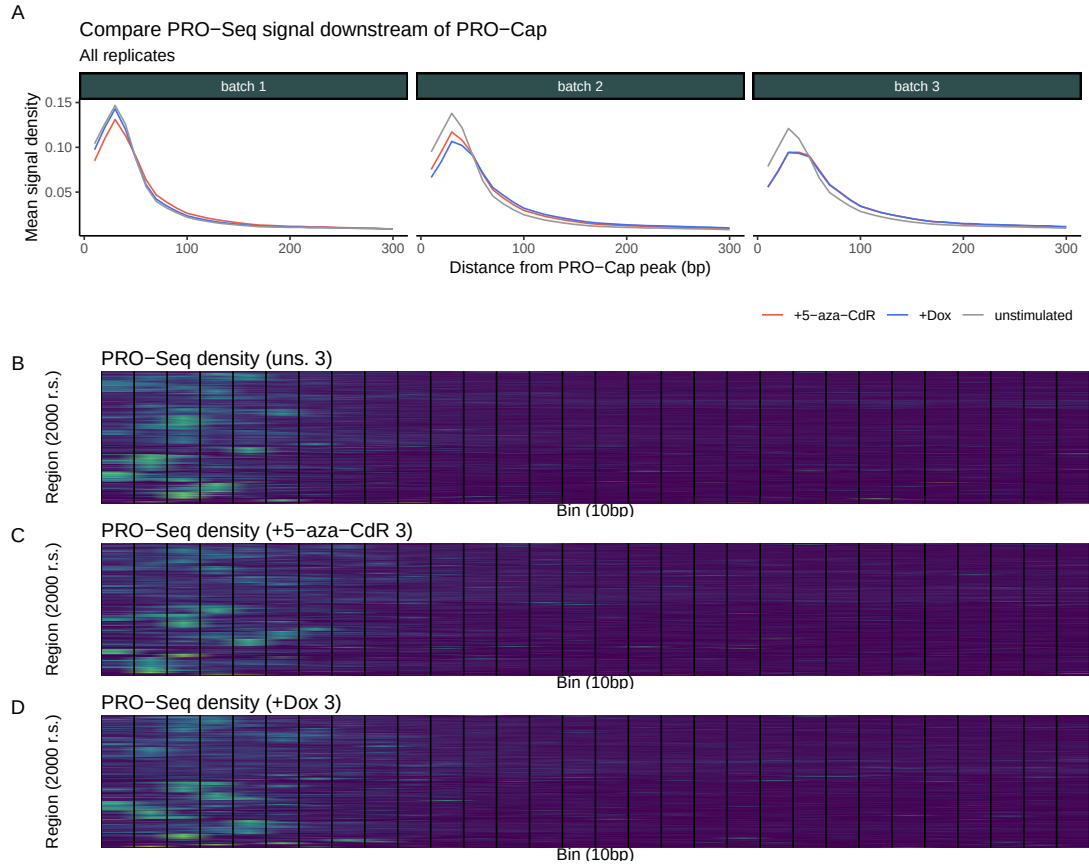


Figure 52: **PRO-Seq read distribution downstream of PRO-Cap peaks compared between conditions.** A. Line plot showing the mean read count density in each 10bp bin downstream of a PRO-Cap peak in each replicate. B-D. Proportion of PRO-Seq signal overlapping each 10bp bin for different conditions in batch 3, regions clustered together using complete linkage and euclidean distance. In each case a random sample (r.s.) of 2000 regions is used.

the contribution from each upstream aaTSS⁵⁹.

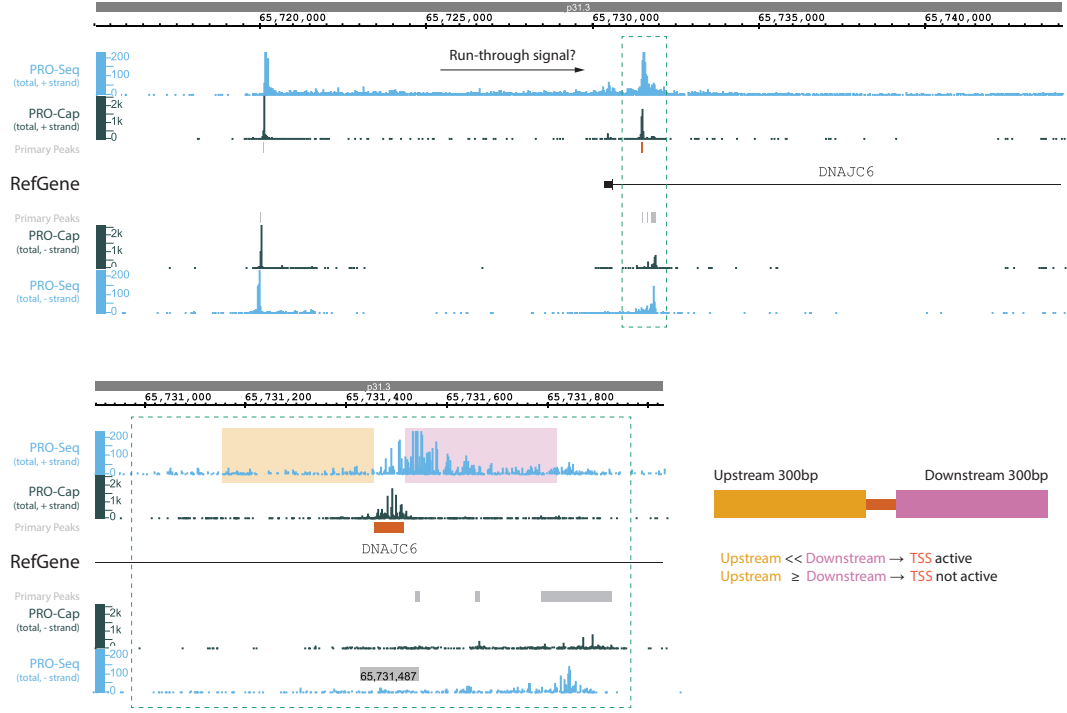


Figure 53: **Example of PRO-Seq run-through.** Genome browser view of PRO-Cap peak whose downstream PRO-Seq signal may be contaminated with nascent fragments which initiated at an upstream TSS. The tracks show total PRO-Cap and PRO-Seq signal for each strand, the merged aaTSS peaks (Primary Peaks), and gene annotation from RefSeq for reference. The 300bp regions flanking the peak which are used to calculate the run-through ratio are highlighted in the zoomed-in view of the peak of interest.

Similarly to Booth et al. [24], we analyse the prevalence of run-through by comparing the PRO-Seq read count around the aaTSS peaks, namely the 300bp regions upstream and downstream of the PRO-Cap peak⁶⁰ (Figure 53). We define the run-through ratio (RTR) for $aaTSS_i$ as:

$$RTR_i = \log_2\left(\frac{down_i + 1}{up_i + 1}\right) \quad (2)$$

where up_i and $down_i$ are the total PRO-Seq signal in the 300bp upstream and

⁵⁹If many aaTSSs are inactive, then we risk double-counting changes if a change in an active TSS propagates through multiple inactive TSSs.

⁶⁰PRO-Cap peaks must be more than 1kb apart to avoid counting TSS clusters.

downstream of $aaTSS_i$, respectively⁶¹. The density of the RTR shows that the majority (92.4%) of TSSs are supplying elongating RNAPs, with $RTR > 0$ ⁶² (Figure 54A). Many of the largest increases are from highly transcribed (measured by reads/bp) regions, suggesting that we aren't overly biased by high variance outliers from low read counts (Figure 54B). Therefore fully inactive aaTSSs are quite rare, and run-through signal constitutes less than 10% of the read count for over 54.7% of aaTSSs.

When comparing the conditions we saw no difference between RTR upon mC and/or hmC perturbation, and average profiles are very consistent (Figures 54C-D). We checked whether the individual aaTSSs also showed similar RTRs to ensure that the density summaries didn't mask any differences, but the strong correlations suggested minimal changes in TSS activity genome-wide (Figures 54E-G). Based on high correlations, we selected two groups of peaks for annotation-based analysis, namely low-RTR and high-RTR peaks, with $RTR < 0$ and $RTR > 6$ for all 3 conditions, respectively. The low-RTR peaks were disproportionately found in intronic and intergenic regions whereas high-RTR peaks were enriched in promoters and 5'UTRs (compared to random sampling), suggesting more productive initiation occurs in coding regions (Figure 54H).

4.2.8.6 Elongation

After defining regions of initiation using the HMM peak caller as above, we wanted to check whether there were any differences in downstream RNAP occupancy.

For each primary peak, $Pr_i = [SS_{Pr_i}, ES_{Pr_i}]$, the fixed X bp elongation region E_i is defined as the X bp downstream of $ES_{Pr_i} + 100$. Only elongation regions which don't overlap with other elongation regions or primary peaks are considered to avoid contamination. As X increased, the number of candidate elongation regions decreased however longer regions were more likely to contain higher read counts and thus have higher power to call differential RNAP occupancy. We also defined variable length elongation regions as follows: the fixed 50kb elongation regions are merged, we then extract the complement between padded primary peaks and this set, and lastly filter for a minimum length of 500bp. The resultant set of regions is the largest of the groups

⁶¹We add 1 to avoid dividing by zero

⁶²In fact the signal at least doubles for almost 80% of peaks.

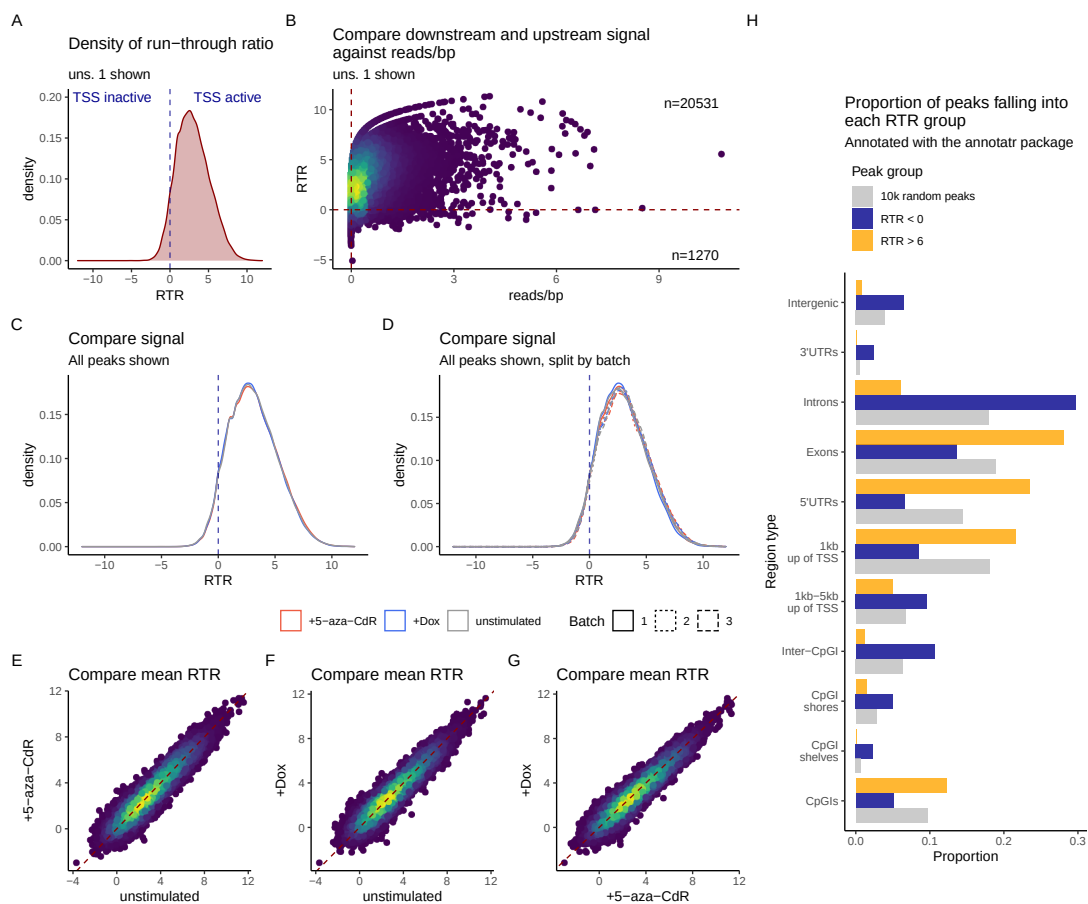


Figure 54: **Run-through ratio (RTR) analysis.** A. Density of RTR values in unstimulated sample 1. B. Density scatter plot comparing the mean read count per bp across the combined upstream and downstream regions around the aaTSS peak. C-D. The RTR ratio densities for each condition separately, and split by batch. E-G. Scatter plots comparing the RTR between conditions, with each point representing a aaTSS peak. For each peak we consider the mean of the RTRs across 3 replicates. H. Bar plots showing the proportion of peaks overlapping each region type, annotated with the basic annotation from `annotatr` [32]. We compare the expected proportions (based on 10k randomly sampled peaks) with peaks that consistently show inactivity ($RTR < 0$), and peak that are highly active in all samples ($RTR > 6$).

and contains regions length 500bp - 50,000bp. See examples of the defined regions in Figure 55.

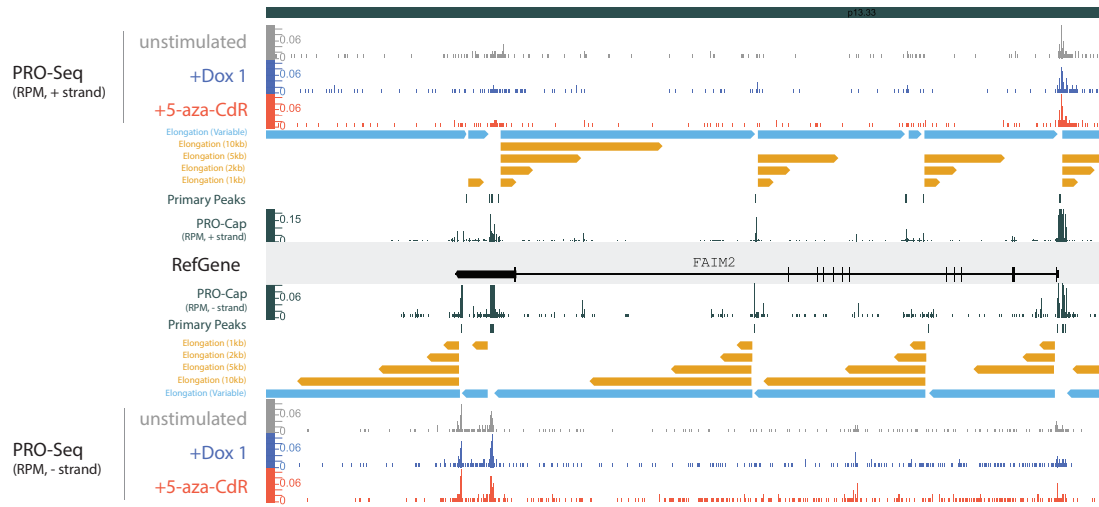


Figure 55: **Annotation Independent Elongation Region examples.** Elongation regions defined using different methods based on HMM-calculated primary peak annotations, with strand-specific PRO-Cap signal (combined across all samples) and PRO-Seq signal (combined across conditions), both CPM normalised.

To identify elongation regions with changes in RNAP occupancy after each treatment we quantified the PRO-Seq signal in the same elongation regions in each replicate and used these as input to DESeq2 [153] (Section 2.4.12). Of the constant-width regions, the 1kb set had the fewest examples of differential signal despite containing the most regions (especially when focussing on high-magnitude changes) suggesting that the counts in these short intervals were too low to definitively detect changes in most cases (Figures 56A and B). The variable length elongation set captured the most differentially occupied elongation regions as it was both the largest and contained the longest regions.

When the width was increased, we progressively captured more changes with consistent observations showing that the highest magnitude changes (those with $|LFC| > 1$) were positive, i.e. +5-aza-CdR and +Dox were more likely to increase RNAP occupancy in elongation regions (Figures 56A and B). Demethylation with +5-aza-CdR had a stronger effect than TET1-CD upregulation, with over three times as many up- and twice as many down-regulated regions in the variable length elongation reference group.

These results suggest that demethylation is more disruptive, but both treatments have the potential to induce changes in RNAP occupancy.

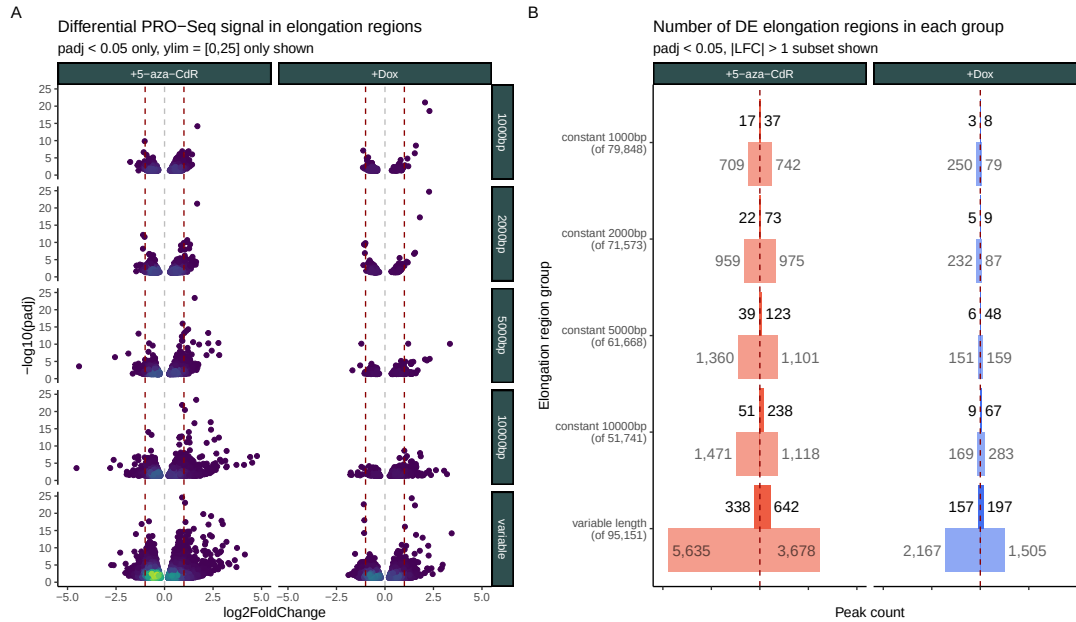


Figure 56: **PRO-Seq differential signal in elongation regions.** A. Volcano plots with each point representing an elongation region with significant differences (p-adjusted ≤ 0.05) in signal between unstimulated and +5-aza-CdR (left) and unstimulated and +Dox (right). To aid visualisation the y-axis has been trimmed to remove outliers. B. Bar plots showing the number of elongation regions with significant up-regulation (right counts) and down-regulation (left counts), with the subset of most altered (defined as those with absolute $\log_2\text{FC} > 1$) highlighted.

4.2.9 Supervised analysis of RNAP initiation and elongation

In the above analyses we've focussed on genome-wide changes independent of where the initiation peaks fall. In this section we would like to quantify:

- How does the annotation compare to the data-driven TSSs?
- How many of the PRO-Cap peaks are TSS-proximal?
- Do any of the TSSs with changes in initiation and/or elongation coincide with annotated gene TSSs?

Initiation Search Re- gion	Notes
[TSS - 30, TSS + 300]	Count reads. [226]
[TSS - 50, TSS + 250]	RPKM normalised. [261]
[TSS - 100, TSS + 400]	RPK at 50bp window with highest read count. [245]
[TSS - 250, TSS + 250]	The position with the maximum count in this 500bp region, at least 4+ reads above background. [24]

Table 4: **Annotation derived TSSs.** Previous approaches to defining transcription start sites using genome annotation.

We will use the GENCODE annotation (Section 2.1.1), containing both gene and transcript annotation to answer these questions and thus isolate gene candidates whose transcription was affected by mC and/or hmC perturbation.

4.2.10 TSS identification

To link differential mRNA expression to changes in active RNAP occupancy and DNA modifications, we would ultimately like to define active TSS(s) and a corresponding elongation region for each active gene.

Many groups have used available gene annotation for this purpose without taking regions outside the promoter into account (see Table 4). Elongation regions can be similarly defined by avoiding promoter-proximal and TES-proximal signal⁶³.

When gene bodies of different genes intersect, assignment of RNAP signal to genes becomes complicated. Various supervised (annotation-based) approaches have attempted to avoid this ambiguity. For example, Nojima et al. [176] removed any padded genes ([TSS-1kb,TES+3kb]) which overlapped, while Sanz et al. [205] considered strand-specific overlaps and defined a “clean” set of genes such that the nearest strand-specific neighbour was at least 15kb downstream of the poly(A) site.

Alternatively, rather than assuming all genes are active, one could incorporate a run-through ratio similarly to Booth et al. [24]. If the peak is considered active (RTR>0),

⁶³For example, in their quantification of expression based on GRO-Seq data, Bi et al. [18] calculated the signal in the first 30kb of the gene bodies, excluding the TSS, [TSS+1000bp, TSS+30kb].

then the gene is active (and has potential to contaminate signal in any overlapping genes). In these cases, the analysis relies heavily on accurate TSS annotation.

By using PRO-Cap data to define data-driven TSSs we can:

- Identify which TSS, of multiple possible TSSs, is used for each active gene. Although TSSs often overlap, using the longest possible transcript in each case may mislabel some TSSs for this cell line as alternative TSS usage is common [237]. In some cases we may see genes with multiple unique active TSSs, in which case the TSS furthest downstream could be incorrectly labelled as pausing signal in less sensitive analyses.
- We can identify gene-body spurious initiation, which can't be done with annotation alone (it is unexpected by definition). Apart from being an interesting phenomenon to analyse in the context of DNA modification changes, small gene-body peaks can contaminate elongation signal and thus wrongfully suggest changes in active elongation.
- Analyse initiation genome-wide. In PRO-Cap, coverage is not limited to genic regions, therefore we can look at other functional regions, e.g. changes in enhancer activity.

Thus, we assigned the data-driven initiation peaks to functional regions for incorporation with the mRNA analyses.

4.2.10.1 Transcript annotation analysis

The problem of assigning aaTSSs to genomic elements is non-trivial due to the complexity of the human genome. The annotation used contains 208,379 transcripts covering 53,160 unique genes. Overlaps between unique genes are awkward to deal with and the level of detail of the annotation is important.

Most of the transcripts (68.4%) in the database are associated with protein coding genes (PCGs) (on average 7.1 transcripts per gene) and long non-coding RNAs (lncRNA) (22%) (Figure 57A). Most non-PCGs were found ~ 100 kb from the closest PCG, however 2,957 lncRNA and 1,205 processed pseudogene transcripts did overlap

with a PCG (Figure 57B-C). In total, this means that 2,827 PCGs (14.2%) overlap a non-PCG element. Before removing almost 3,000 genes from the analysis we decided to check the overlaps and found that many of the overlapping non-PCGs didn't have a high support level.

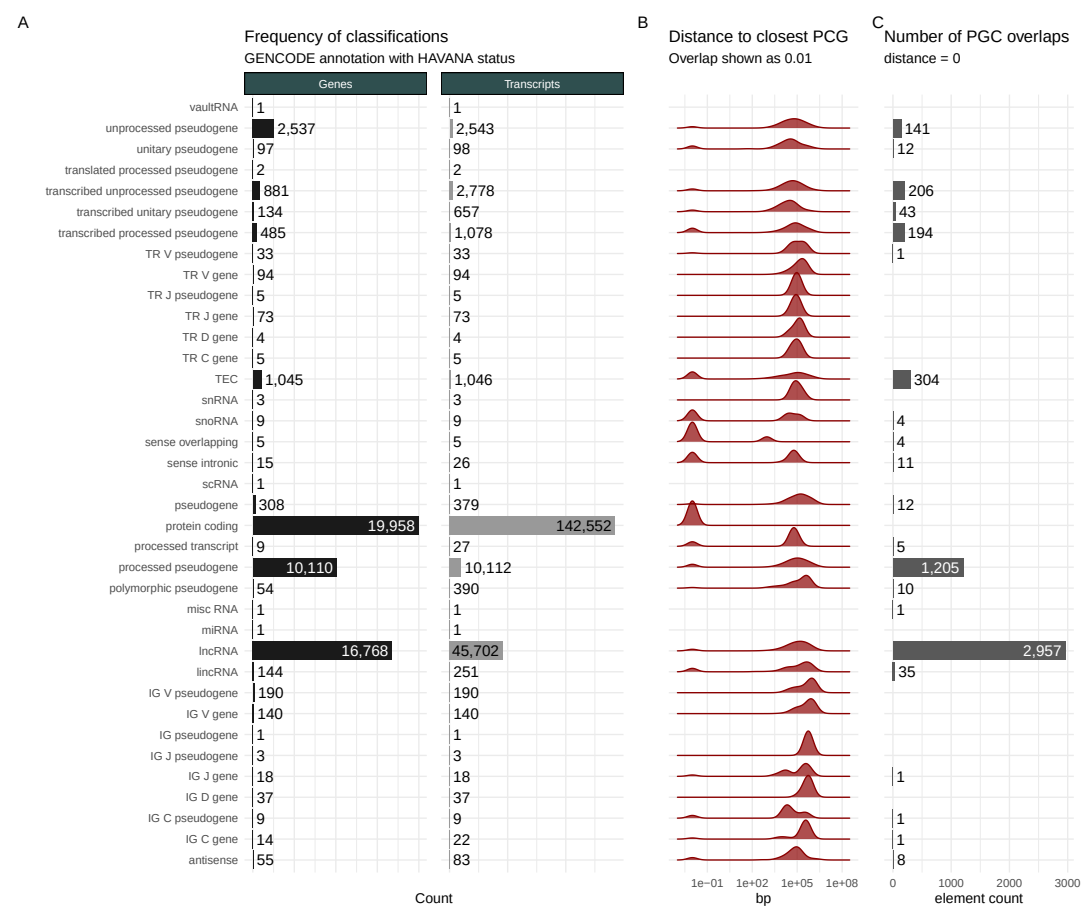


Figure 57: **Analysis of GENCODE annotation.** A. Bar plots showing frequencies of genes and transcripts falling into each gene category. B. Density plots of distances between each transcript and the nearest protein-coding gene (PCG) on the same strand. C. Bar plot showing the number of unique transcripts that overlap a PCG.

Not all transcript annotations are equally supported, and this is incorporated into the `gtf` files through the `transcript_support_level` column⁶⁴. In the annotation used, a substantial proportion of transcripts fall into the lower support levels (Figure 58A). By looking at the categories directly we can see that there are many high-support lncRNAs

⁶⁴See https://m.ensembl.org/info/genome/genebuild/transcript_quality_tags.html

and pseudogenes, but most of the high confidence transcripts are PCGs (Figure 58C). We want to avoid overlapping genes in our annotation-driven analyses, however it's clear that PCG annotations have varying levels of support and thus we may over-filter by considering all PCG transcripts. We calculated the frequency of pairwise overlaps between PCGs and found that although the high-support did overlap some low-support transcripts, the counts weren't substantial (Figure 58B). That said, there were 431 genes with overlaps and these mostly fell into one of three groups:

- AC/AL/AP transcripts which are novel
- combination transcripts, e.g. ATP6V1G2-DDX39B
- other

This analysis showed that removing low-support transcripts would not significantly increase the number of non-overlapping genes, therefore we kept all transcripts in the annotation, regardless of TSL.

4.2.10.2 Defining initiation regions

In previous analyses we called peaks on each replicate separately in consideration of replicate variability, however calling peaks on the combined PRO-Cap signal tracks had the benefit of identifying regions with lower (but consistent) transcription initiation⁶⁵. We called HMM peaks (in a strand-specific manner) on the combined tracks and identified 148,405 peaks. After trimming the set contained 216,008 primary peaks. Some untrimmed HMM peaks contained more than a single primary peak but in most cases a single primary peak dominated each raw HMM peak based on counts (Figures 59A and B).

The annotation contains over 208k unique transcripts, with 191,662 unique TSS-gene matches. We systematically assign each primary peak to genes based on overlaps with gene annotation (Table 5). Using this overlapping system we can hierarchically annotate peaks based on their overlaps with:

⁶⁵If a replicate has 4 reads in a 10bp window, this window is unlikely to be called as a peak. However, if the 9 samples show similar transcription then the window will have 36 reads which be considered as enriched based on the threshold of 10 reads.

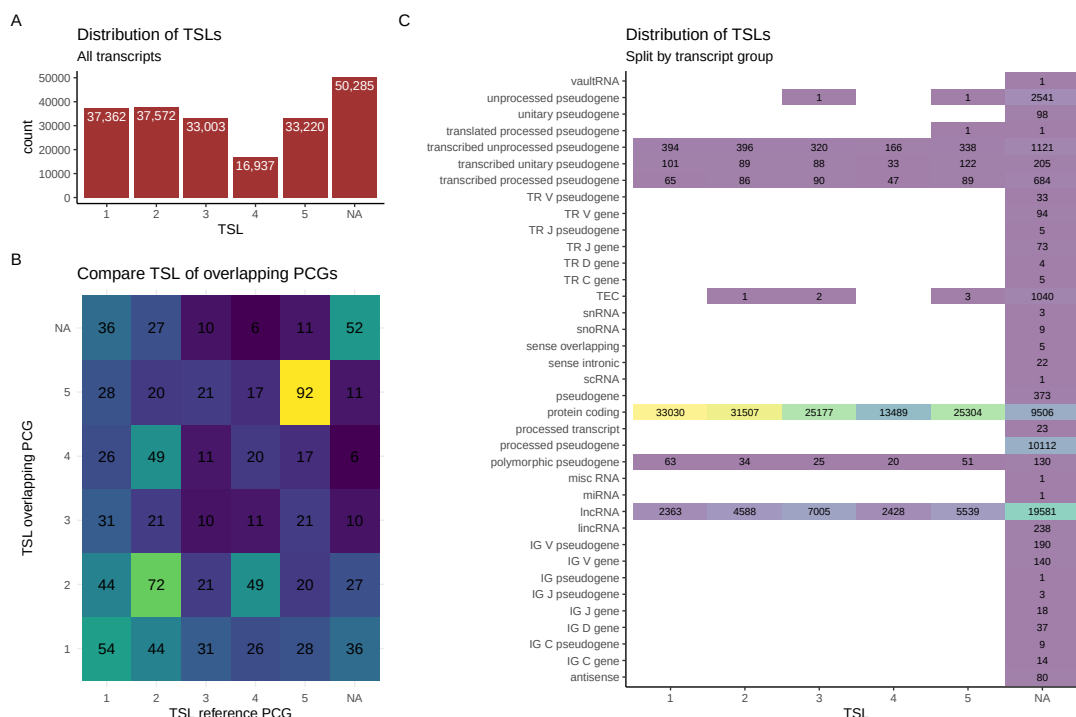


Figure 58: **Analysis of transcript support levels (TSLs).** A. Bar plot showing the counts of each TSL in the annotation. B. Frequency of TSL combinations between overlapping protein coding genes (PCGs). For example, if a PCG with TSL 1 overlaps a gene with TSL 5 then the count at (1,5) and (5,1) will increase by 1. C. The number of transcripts falling into each TSL per gene category. In total 431 PCGs showed overlaps for at least one transcript.

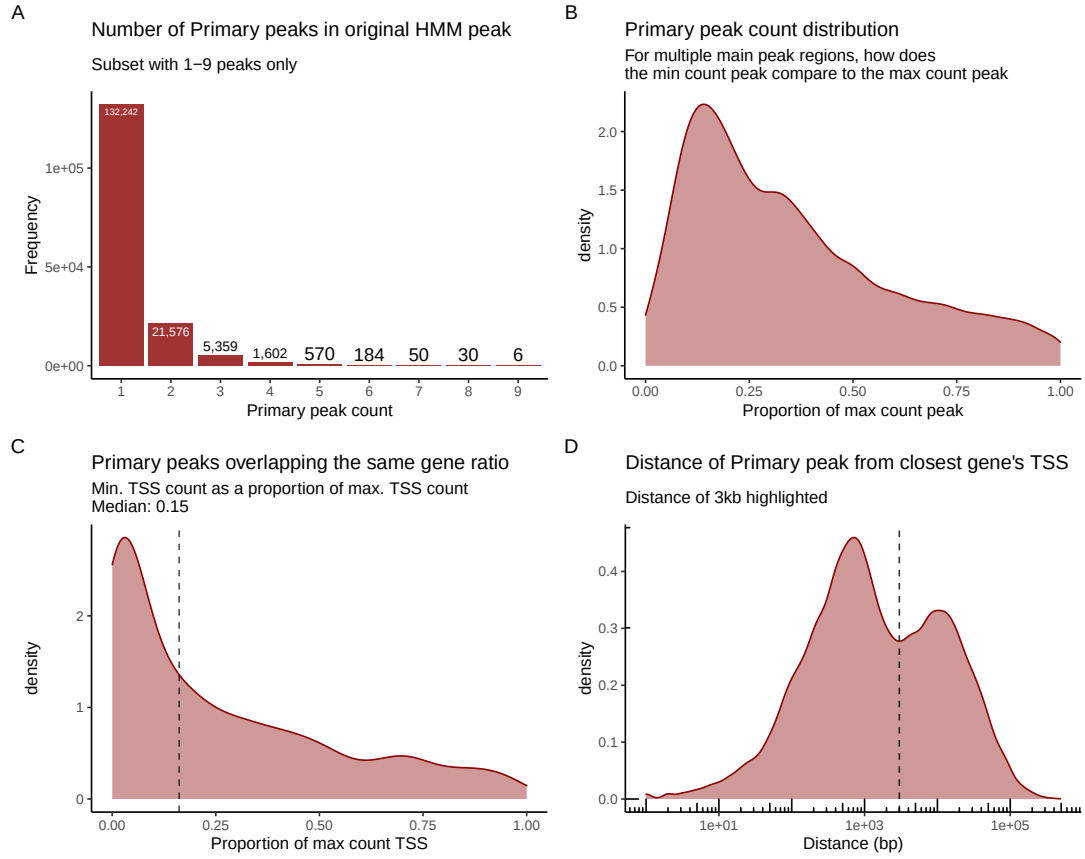


Figure 59: **Primary peak distribution.** A. Bar plot showing the number of untrimmed HMM peaks with 1-9 primary peaks after trimming. B. For HMM peaks with more than one primary peak, how does the count (n_{max}) in the major peak (the peak with the highest count) compare to the count (n_{min}) in the lowest count peak. The plot shows the density of the proportion n_{min}/n_{max} for each HMM peak.

- TSS - annotated TSSs
- TSS (AS) - annotated TSSs with strand flipped
- 1kb TSS - annotated TSSs padded by 1kb on each side⁶⁶, i.e. peak is within 1kb of a TSS
- 1kb TSS (AS) - annotated TSSs with strand flipped padded by 1kb on each side
- GB - full transcript, i.e. starting at TSS and ending at TES
- GB (AS) - full transcript with strand flipped

TSS	TSS (AS)	1kb TSS	1kb TSS (AS)	GB	GB (AS)	Peak group
Y	N					TSS peak
N	Y					TSS (AS) peak
Y	Y					TSS strand overlap
N	N	Y	N			1kb TSS peak
N	N	N	Y			1kb TSS (AS) peak
N	N	Y	Y			1kb TSS strand overlap
N	N	N	N	Y	N	GB peak
N	N	N	N	N	Y	GB (AS) peak
N	N	N	N	Y	Y	GB strand overlap

Table 5: **Peak classification requirements.** For each PRO-Cap, we check for overlaps with GENCODE annotated TSSs (TSS), TSSs padded by 1kb on both sides (1kb TSS), and the full transcript (GB)

where the strand information matches. For the antisense (AS) groups we flip the strand before calculating overlaps. For example, a 1kb TSS (AS) peak must not overlap an annotated TSS directly (on either strand) but must be within 1kb of a TSS on the opposite strand only.

Using this system we can differentiate between initiation sites that overlapped with GENCODE annotated (TSS peaks), those that are close (1kb TSS) and those that overlap within the gene body but not close to annotated TSSs (GB). Ambiguities based on strand are accounted for by “strand overlap”, e.g. if a peak is within 1kb of a TSS on each strand (but doesn’t directly overlap either) then it is both a 1kb TSS and 1kb TSS (AS) peak, cannot be unambiguously assigned to a gene and is classed as 1kb TSS strand overlap. We found that, of ~202k peaks, only ~15,528 overlapped with an annotated TSS on the sense strand with 26,355 peaks within 1kb of annotations and a further

⁶⁶Based on the metagene profiles we saw that most PRO-Cap signal is found near annotated TSSs, therefore we expanded our search area to 1kb around annotated TSSs.

26,239 overlapping gene bodies away from TSSs (Figure 60). The overlapping peaks made up small proportions of the TSS and GB peaks (1.07% and 7.03%, respectively), but the frequency of TSS pad overlaps resulted in a substantial 1kb TSS strand overlap proportion (21.27%). These are cases where genes diverge, and thus it's difficult to classify peak signal into sense or antisense. The prevalence of antisense initiation both near TSSs and in gene bodies was intriguing and will be discussed in the context of methylation changes in the next section.

The “peak group” classification was refined based on further overlaps. A peak can be a TSS peak overlapping more than one unique gene, and thus cannot be assigned unambiguously. Therefore, for each peak group without strand overlaps, we check how many unique genes fulfil the criteria of the peak⁶⁷. Although most peaks (96.4%) could be assigned to a single unique gene, there were some overlaps (Figure 60) which we split into:

- Subgenes (0.49%) - combination genes e.g. peak overlapping PMF1 and PMF1-BGLAP
- New gene annotation (2.48%) - alternative genes (those with an “A[CDFLP][0-9]” prefix), e.g AC118553.2 is an alternative transcript of SLC35A3
- 4+ overlaps (0.03%) - Cases where there are four or more unique gene names fulfilling the criteria
- True overlap (0.67%) - all other overlapping groups

We chose to filter out the “4+ overlap” and “True overlap” peaks (990 in total) due to the ambiguity in annotation, but keep the other overlap types, leaving 140,158 genic peaks and 20,278 unique genes⁶⁸.

We calculated the number of peaks annotated for each gene (sense and antisense separately) and found that only 5,874 had a single sense peak⁶⁹ (Figure 61A). Of the 19,005

⁶⁷If a peak overlaps with a TSS of gene A which is found in the gene body of gene B then it will still be assigned to gene A only due to the hierarchy.

⁶⁸For the subgene peaks, the peak is assigned to each of the subgenes and is thus counted more than once (same for the alternative genes).

⁶⁹Curiously there are 1,272 genes with no sense peak but some antisense peaks.

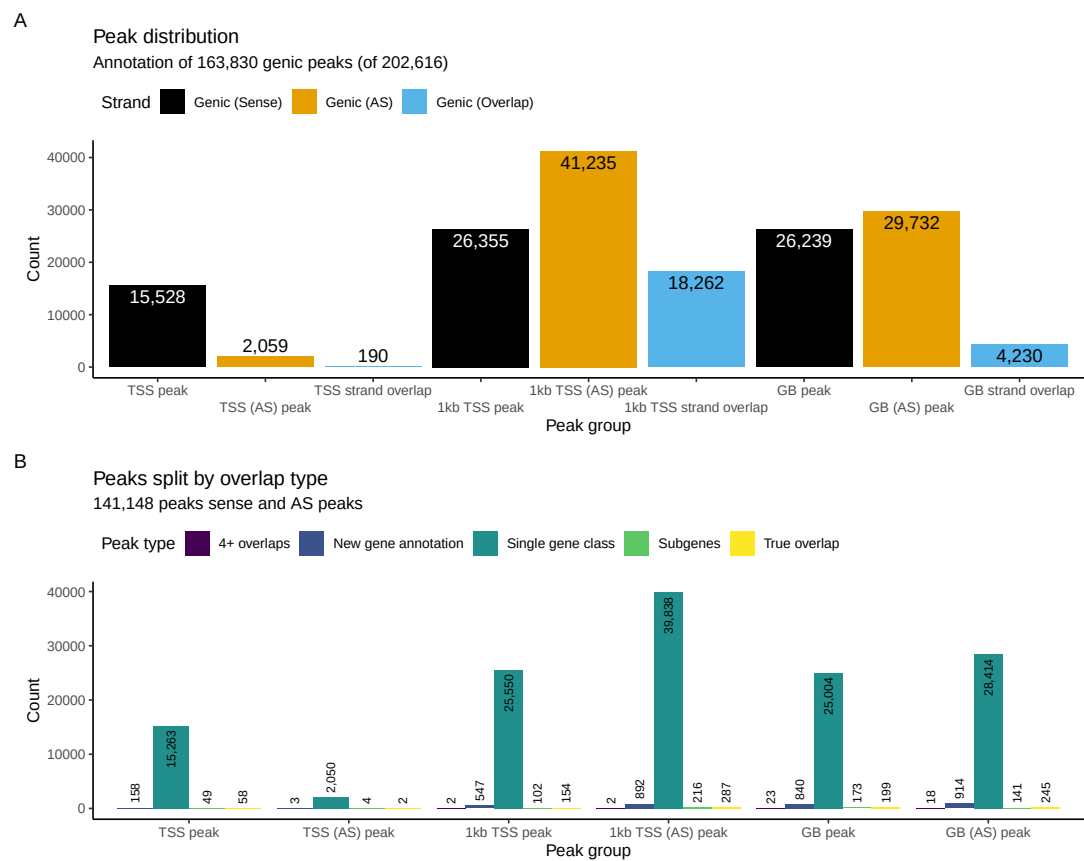


Figure 60: **Peak annotation split.** A. Bar plot showing the number of PRO-Cap trimmed peaks falling into each peak group based on the classification in Table 5. B. Bar plot showing the number of PRO-Cap trimmed peaks in each overlap category.

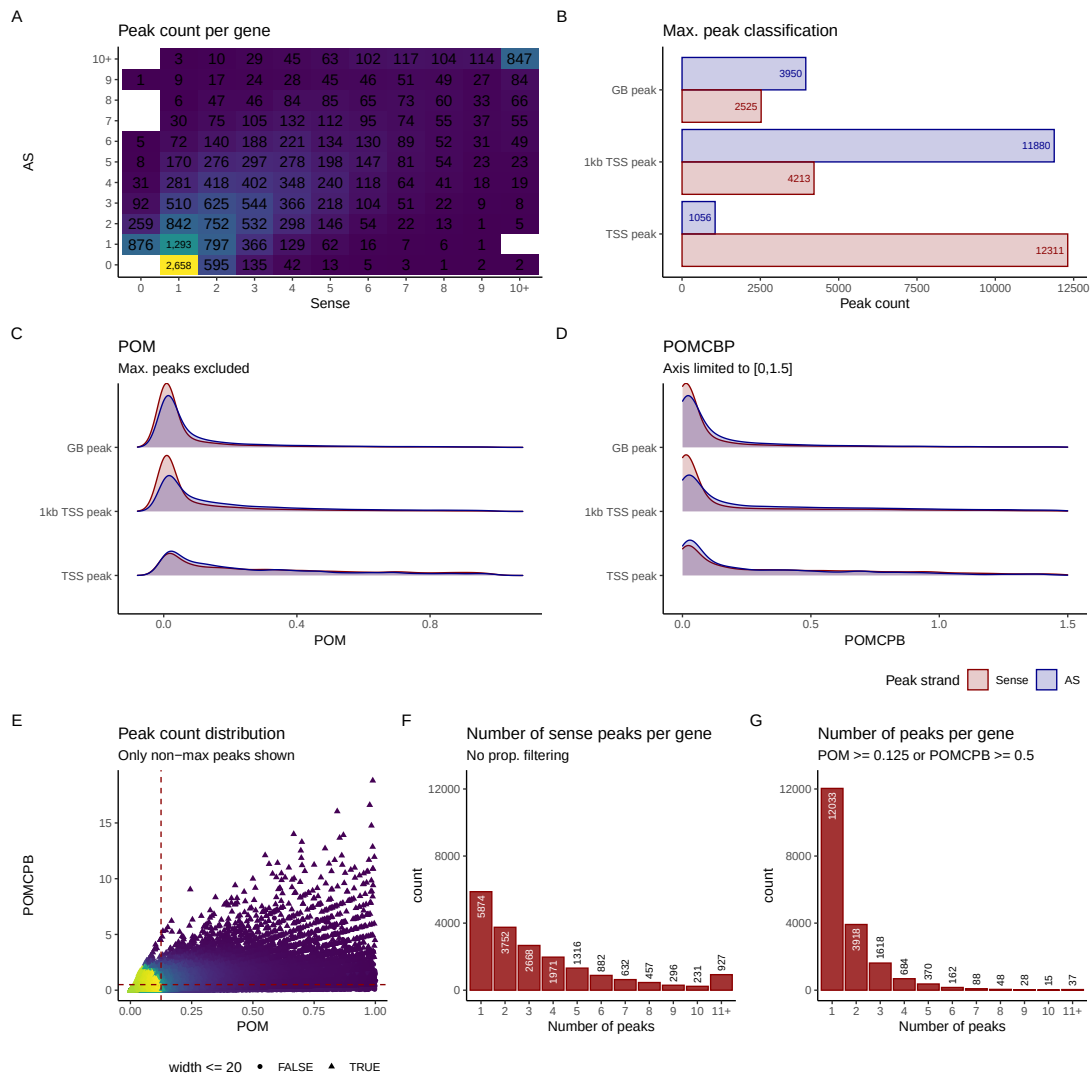


Figure 61: Gene peak filtering. A. Bar plots showing the number of max peaks A-B. Histograms showing the distribution of proportion of maximum (POM) and proportion of maximum count per base pair (POMCBP) values calculated from genes with multiple TSSs (the maximum peak of each gene is left out). C. Density scatterplot showing how the POM and POMCBP metrics are related. The shape of the points is determined by the width of the peak, and the thresholds used are included as dashed lines. The plot is coloured based on density (highest density region is in yellow). D. Bar plot counting the number of genes with 1,2,...,20 peaks. E-F. The effect of proportion filtering on the number of peaks per gene ($\text{POM} \geq 0.125$ in both, and $\text{POMCBP} \geq 0.5$ or 1.)

genes with a sense peak, 13,131 (69%) showed evidence of transcription from multiple active TSSs (see examples in Figure 62A and B) however, based on the genome browser, these TSSs weren't equally active and many multiple-peak genes showed evidence of one major PRO-Cap peak with multiple weaker initiation regions. We decided to quantify this observation as follows:

- For each gene j with multiple peaks, we determine the major peak p_{m_j} with the highest count m_j .

$$m_j = \max_k [n_{p_{k,j}}] \quad p_{m_j} = \{p_{i,j} | n_{p_{i,j}} = m_j\} \quad (3)$$

- For each peak $p_{i,j}$ in gene j we calculate
 - the proportion of maximum (POM) as the ratio of the number of reads in $p_{i,j}$ (denoted $n_{p_{i,j}}$), and m_j .

$$POM_{p_{i,j}} = \frac{n_{p_{i,j}}}{m_j} \quad (4)$$

- the proportion of maximum per base ($POMCBP_{p_{i,j}}$), i.e. the POM normalised by the lengths of the peak, $|p_{i,j}|$ and $|p_{m_j}|$.

$$POMCBP_{p_{i,j}} = \frac{n_{p_{i,j}}/|p_{i,j}|}{m_j/|p_{m_j}|} \quad (5)$$

Most of the sense major peaks p_{m_j} overlapped with annotated TSSs directly but antisense signal was generally found close (but not overlapping the TSS) (Figure 61B). This suggests that the annotation was the optimal TSS to use for 64.6% of genes with an active peak, but the proportion of major peaks falling into gene body regions (13.3%) is not negligible⁷⁰. Thus using PRO-Cap initiation signal to define initiation, pausing, and elongation regions is more accurate for over 35% of genes.

Genes with multiple peaks tended to have one major peak since the distributions of POM and POMCPB were highly right skewed (Figure 61C and D). The TSS-overlapping

⁷⁰These peaks may be accounted for by enhancers.

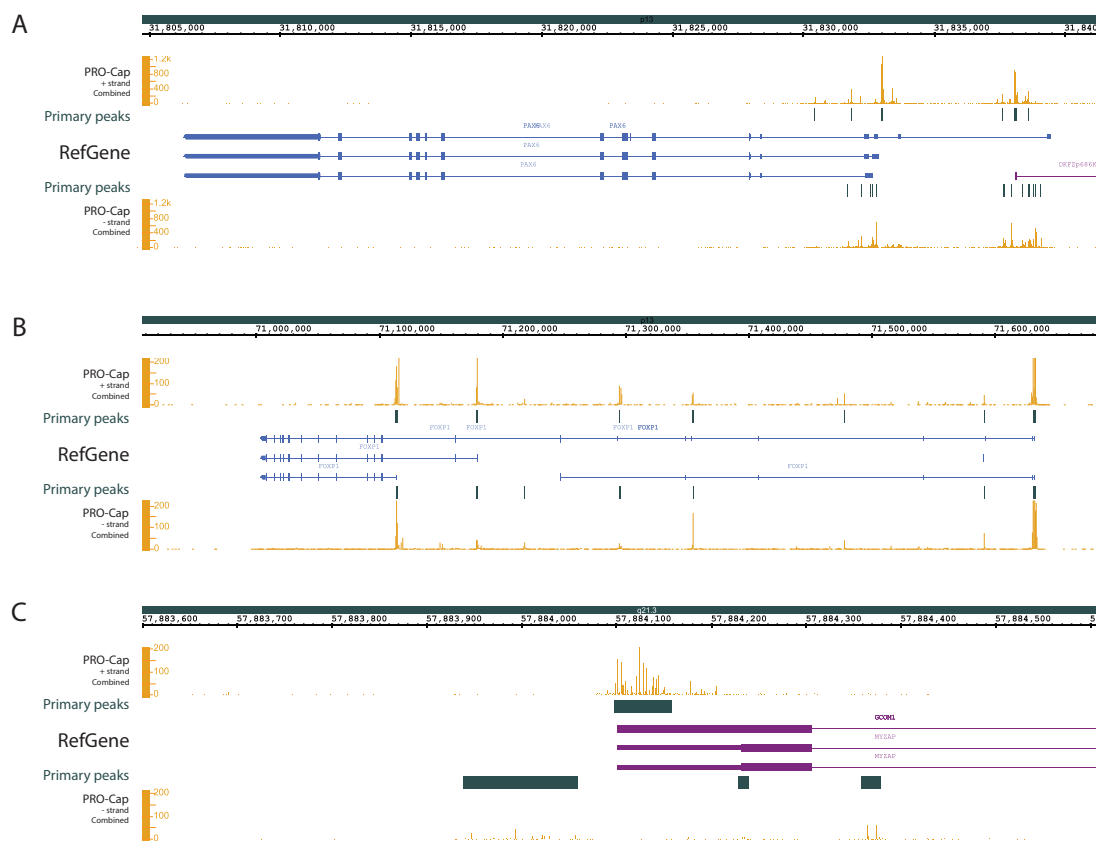


Figure 62: Examples of peak-TSS overlaps. Genome browser tracks showing the combined (strand-specific) PRO-Cap signal, the primary peaks called on each strand, and a reference annotation for 3 genes (A. PAX6, B. FOXP1, C. MYZAP and GCOM1). The orientation of the gene is shown as the colour, purple for plus strand and blue for minus strand.

peaks (both sense and antisense) were slightly more likely to be similar in magnitude to the major peak⁷¹, and antisense peaks had a less sharp distribution. Most of the outliers were short (10-20bp) (Figure 61E) and, based on the combined distributions, we chose to filter out peaks with $POM \geq 0.125$ and $POMCPB \geq 0.5$. This increased the number of genes with a single primary peak from 5,874 to 12,033⁷² (Figures 61F-G).

The 6,968 genes which retained multiple initiation regions often showed clusters of PRO-Cap peaks and, by calculating the distances between peaks (end to start) in multiple-TSS genes, we found that 57% fall within a 1kb window⁷³ (Figure 63A). This suggested that initiation in these regions could be clustered into a broad initiation region⁷⁴.

Based on the metagene data, we expected to see an abundance of antisense signal upstream of the TSSs. Indeed, when we calculated the distances between major peaks and any antisense peaks (annotated to that gene) we found that most AS transcription occurs upstream of initiation (Figure 63). The abundance upstream of the TSS is consistent with a model whereby TSS-proximal transcription is bi-directional, but in most cases the RNAPs on opposite strands start transcription from distinct promoters. This is in agreement with [53] who described a model where divergent transcripts arise from distinct promoters flanking the nucleosome-free region.

In this analysis of initiation peak distribution we have filtered some obvious overlap candidates, found that most peaks do not overlap with annotated TSSs, however after proportion filtering we saw that most major peaks do indeed overlap with GENCODE annotated TSSs. The non-major peaks are still largely genic. Initiation is often found on both strands, in multiple loci, but a single major TSS (and AS TSS) often dominates. This allows us to define elongation regions by simply considering any PRO-Seq signal downstream of the most downstream major peak.

⁷¹Multiple similarly active, annotated TSSs.

⁷²More stringent POMCPB filtering resulted in marginally more single peak genes, 12,352.

⁷³65% within 2kb

⁷⁴Thus defining an elongation region downstream of the most 3' peak would be sensible for over half of the multiple-TSS genes (without extending too far into the gene body).

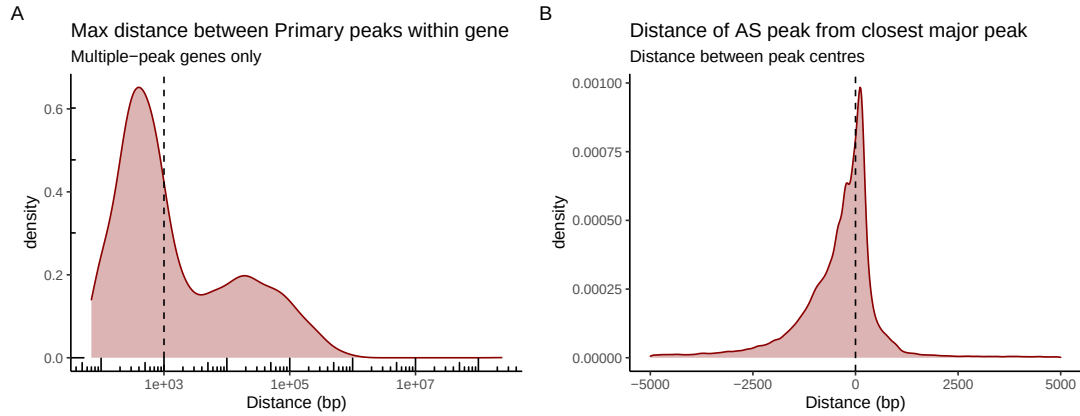


Figure 63: **Peak dispersion.** A. Density of the distances between primary peaks within each gene. Distance of 1kb highlighted. B. Density of distances between each antisense (AS) peak and the closest primary peak within each gene, with negative values for upstream and positive for downstream. In each case the distance is calculated as the distance between peak centres.

4.2.10.3 Definition of elongation regions

This section focusses on how we defined elongation regions for each gene using the results of the initiation analysis. We first group each gene's major peaks, and only consider signal downstream of the most 3' major peak⁷⁵. The elongation signal varies in length⁷⁶ and usually makes up over 90% of the gene body (Figures 64A and B) however we did note the prevalence of pausing near TESs in the metagenes and further analysis of the 5kb upstream of TESs confirmed a slight increase in signal (Figure 64C and D). To avoid the TES-proximal pause, we remove the penultimate 3kb of the gene body and keep only the elongation regions of length 1kb+ for further analysis, this results in 14,334 elongation regions with sufficient length (of which 11,778 correspond to protein-coding genes).

⁷⁵We include an extra 100bp pad to account for pausing. Based on the dedicated pausing analysis this is likely to be slightly too stringent but we wanted to err on the side of caution since pausing signal usually dominates the PRO-Seq data as evident from the metagene plots.

⁷⁶We found that 420 genes actually ended up with negative elongation length (the true TSS was close or downstream of the TES), and the length was exceeded in 1,096 (the true TSS was upstream of the annotated).

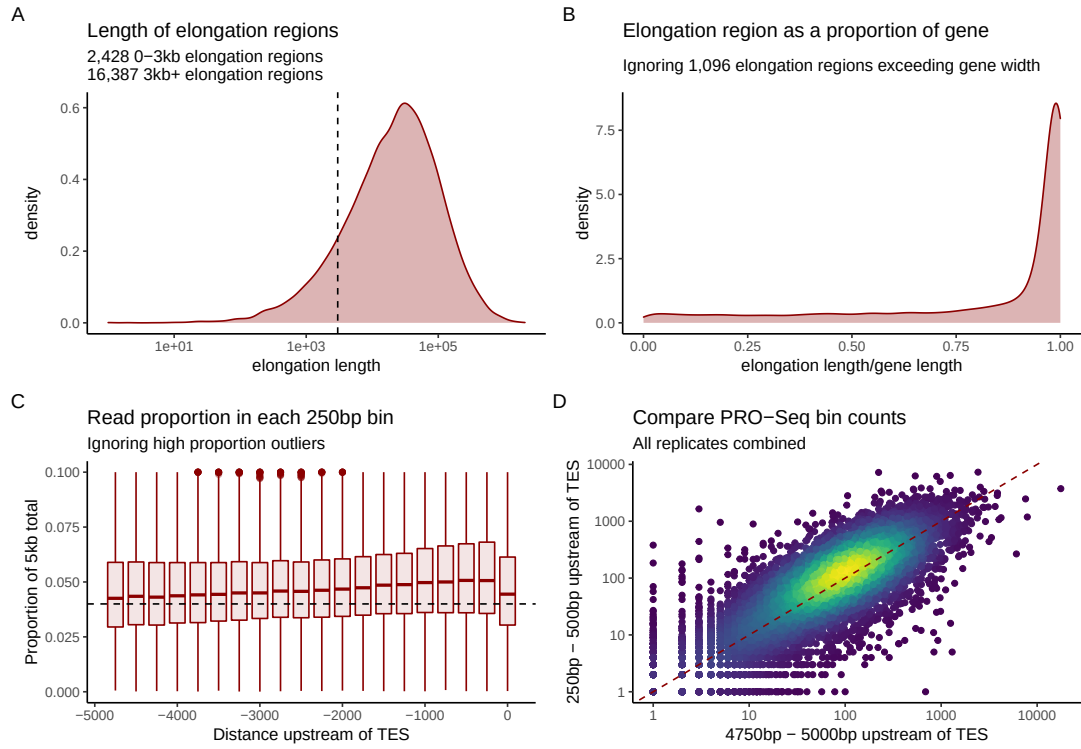


Figure 64: **Supervised elongation regions.** A. Density of the elongation region lengths, with 3kb highlighted. B. Density showing the proportion of the gene body ([TSS, TES]) which the elongation region takes up. C. Boxplots showing the signal distribution in each 250bp bin upstream of the TES. For each gene, we calculate the reads in each bin as a proportion of the total read count in the 5kb region. D. Density scatter plot showing direct comparison of signal in the 250bp regions 5kb and 500bp upstream of the TES.

4.2.11 Effects of DNA modification

With the regions of interest defined based on combined PRO-Cap signal, we break the data back into replicates to calculate the effects of both treatments. The goal is to ascertain which genes are likely to have been affected by the changes in DNA modifications based on the RNAP initiation and occupancy data.

- Based on the annotation-agnostic analysis we saw a prevalence of high-magnitude gains in initiation after both treatments, and a preliminary analysis showed that they were usually genic. Using the annotation, we would like to know
 - Where do these gains occur, i.e. are they TSS-proximal or is there evidence of gains in gene-body TSSs?
 - Do the strands (sense and antisense) behave similarly?
- Elongation is a better predictor of final mRNA levels than initiation and thus we want to know how many genes showed detectable changes in elongation to later contrast those changes with mRNA levels.
- Were changes in mC and hmC related to differences in other stages of transcription? In particular do any genes show changes in pausing, termination, or stalling?

4.2.11.1 Treatment-induced changes in initiation

4.2.11.1.1 Is PRO-Cap representative of initiation frequency?

We would like to begin with a disclaimer, and note that PRO-Cap data is not directly representative of recruitment or initiation frequency due to some technical limitations.

In PRO-Cap we sequence fragments containing both a biotin and 5'-cap, so, in order for a fragment to be included the RNAP must be transcribing (either not paused yet or released from pausing), incorporate a biotin, contain a 5'-cap, and the distance from the TSS must exceed 20bp and be uniquely mapped. Thus any short fragments or those that haven't been capped yet will not be included. This limitation can be addressed by other techniques such as RNAP-ChIP.

Fragmentation is not a step in PRO-Cap, thus we can end up with some very long fragments. Due to lack of sonication step, these tend to be left out, resulting in under-representation of these RNAPs at the TSS in our data (Figure 65). This means that RNAPs that escape pausing faster are more likely to be left out of the PRO-Cap initiation data. However, based on the adapter content distributions we saw that on average 65% of reads included some adapter sequence, meaning that most of the reads were indeed short. More importantly, analyses of PRO-Seq metagene distributions suggested that the majority of signal is indeed promoter-proximal, therefore we do not expect this to affect the data hugely.

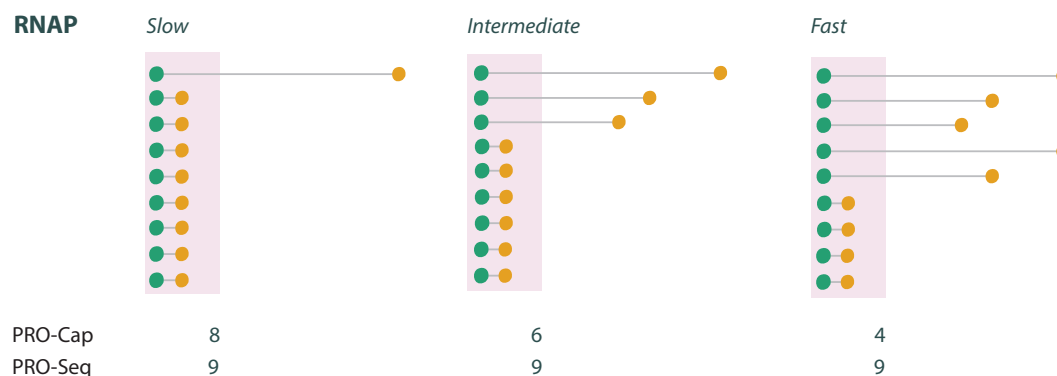


Figure 65: **PRO-Cap long read limitations.** We show a schematic of 3 cases with same recruitment, but differences in the number of reads that contribute to the PRO-Cap signal due to variation in RNAP pause release. Assuming that only short reads will be included in the final data, we can see that whenever the RNAP escapes pausing quickly we will underestimate the number of RNAPs that initiated at that TSS.

4.2.11.1.2 Treatment-induced changes in PRO-Cap signal

To identify differences in PRO-Cap signal between conditions we compared the counts in each peak between the unstimulated and the treated samples using DESeq2 [153] (Section 2.4.12). Of the 167,395 peaks that fulfilled the 10 read minimum, 4.36% (7,306) and 3.04% (5,093) showed significant ($p\text{-adj} < 0.5$) changes in PRO-Cap signal after +5-aza-CdR and +Dox treatment, respectively. The directionality of the change was somewhat unexpected, as +5-aza-CdR induced demethylation resulted in a reduction in

77% of peaks, and TET1-CD upregulation through Dox induction resulted in drops in 64% of peaks (Figures 66A and B). However, the highest magnitude changes were biased toward being positive in each case and most of the peaks with decreases in signal had low fold changes, $-1 < LFC < 0$. Although most peaks with greatest gains had relatively low expression (as estimated by baseMean), the bias toward higher magnitude gains was clear and changes were consistent between treatments, i.e. peaks that significantly changed after both treatments both gained or lost (Figure 66C).

Previous analyses of PRO and GRO data have used a variety of thresholds to define significant changes after DESeq2 analysis (see Table 6). We noted the inclusion of LFC filtering in multiple studies, and chose to add a $|LFC| > 1$ filter to our significance criteria since low magnitude changes weren't very convincing in the genome browser and may have been a technical artefact (Section 4.2.11.1.1). Filtering resulted in a flip in interpretation, in particular demethylation and hmC induction generally upregulate transcription initiation (Figure 66D, see examples in Figure 67). There was substantial overlap between the peaks as expected based on demethylation caused by TET1-CD but some peaks were indeed +Dox specific, suggesting that the changes may be hmC-related.

p-value	p-adj	LFC	Papers
0.001			[159, 158]
0.05		1.25	[196, 246]
	0.01		[24]
0.001		1.25	[245]
	0.0001	1.5	[60]
0.01		2	[17]

Table 6: **Methods to define significant changes from PRO-Seq data.**

To test whether the changes seen are indeed a consequence of changes in DNA modifications, we calculated the mean mC and hmC levels for each initiation peak based on TAPS and TAPS- β data⁷⁷. First, we checked the mC levels of the initiation sites which showed changes after both treatments and found that these initiation regions were more likely to be hypermethylated in the unstimulated sample (Figure 68A and C). A comparison of the mC loss after +5-aza-CdR and hmC gain after +Dox with the

⁷⁷For the 107,532 peaks with covered CpGs, the median and mean CpG count are 4 and 6.9, respectively.

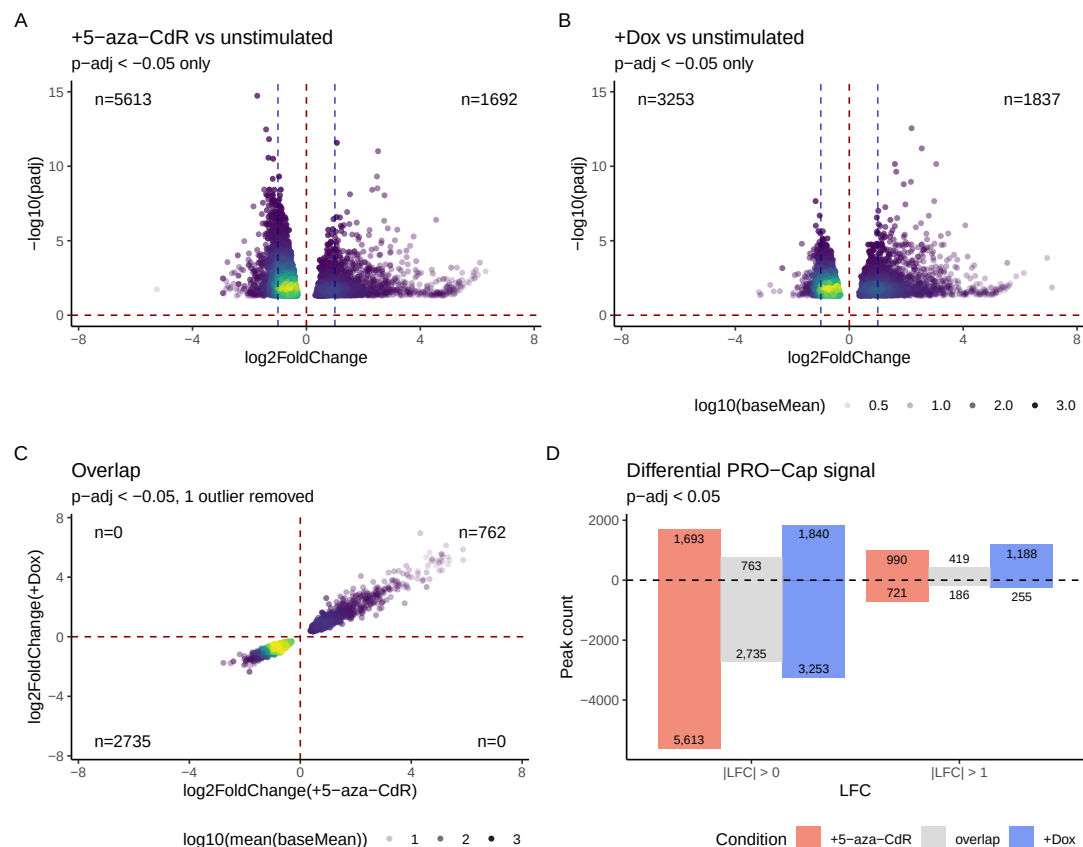


Figure 66: **Differential signal analysis of all trimmed PRO-Cap peaks.** A-B. Volcano plots comparing +5-aza-CdR vs unstimulated and +Dox vs unstimulated, respectively. The density scatter plots show the log2FoldChange and -log10(adjusted p-value) for each peak with a significant (< 0.05) adjusted p-value, with some outliers removed through limiting the x- and y-axis for easier comparison. The counts in each quadrant are included, and dashed lines highlight $LFC = 1$ and $LFC = -1$. C. Density scatter plot comparing the log2FoldChange after each treatment for peaks that have $p\text{-adj} < 0.5$ in both. D. Bar plot counting the number of peaks that are upregulated (positive values) and downregulated (negative values) after each treatment subject to LFC filtering, with the overlapping peak set shown in gray.

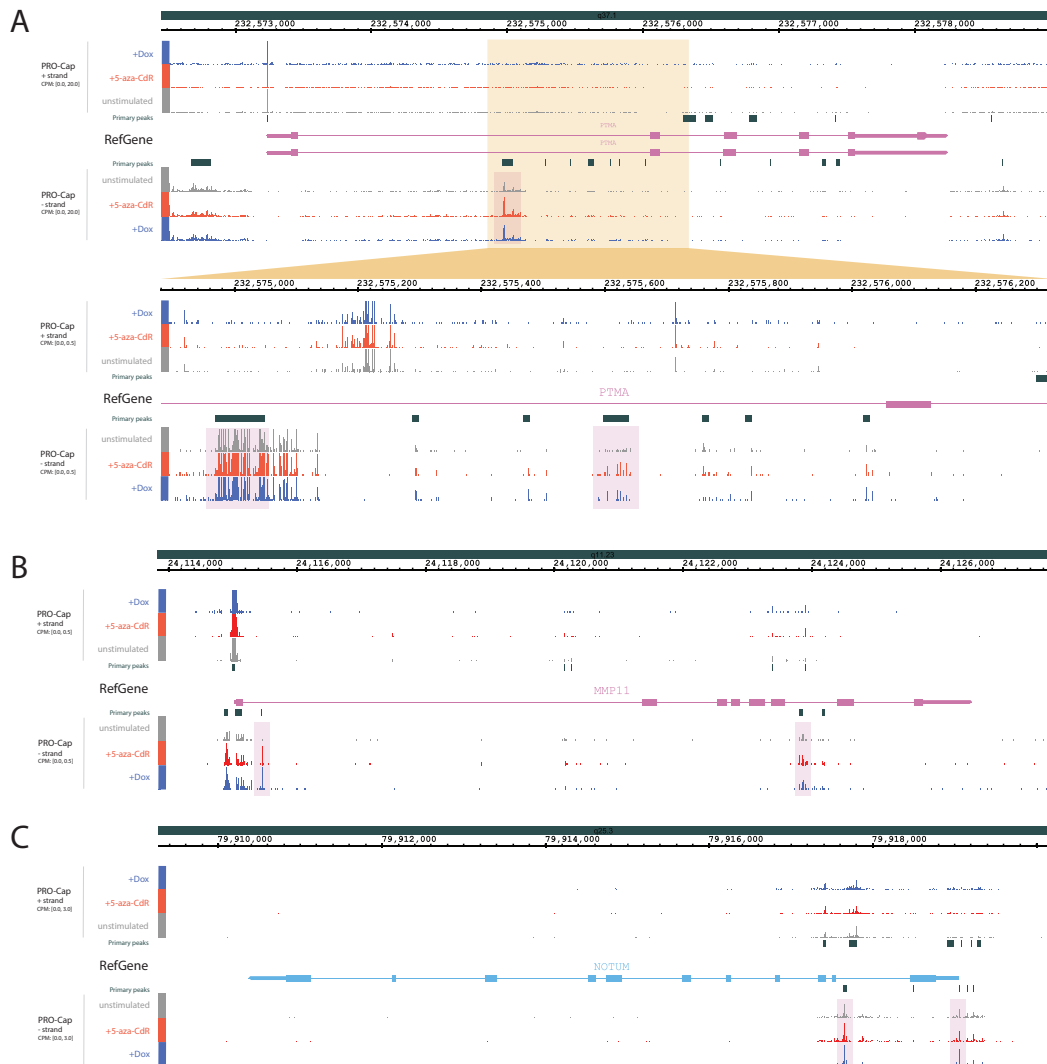


Figure 67: **Examples of peaks with changes in PRO-Cap signal.** Genome browser tracks showing the combined (strand-specific) PRO-Cap signal, the primary peaks called on each strand, and a reference annotation for 3 genes (A. PTMA, B. MMP11, C. NOTUM). The orientation of the gene is shown as the colour, purple for plus strand and blue for minus strand and the peaks with changes in signal are highlighted in each case.

degree of increase (or decrease) in PRO-Cap showed that the most altered peaks were more likely to lose mC and gain hmC compared to those that showed no change (Figure 68B and D), but which is more important? By running a regression to predict log2FC, the coefficients for hmC gain and C gain were 2.32 and 3.58, respectively, suggesting that the level of C gain (full demethylation) may be more important⁷⁸. This is consistent with the overall observation that demethylation with +5-aza-CdR is more disruptive (compared to oxidation induced by +Dox).

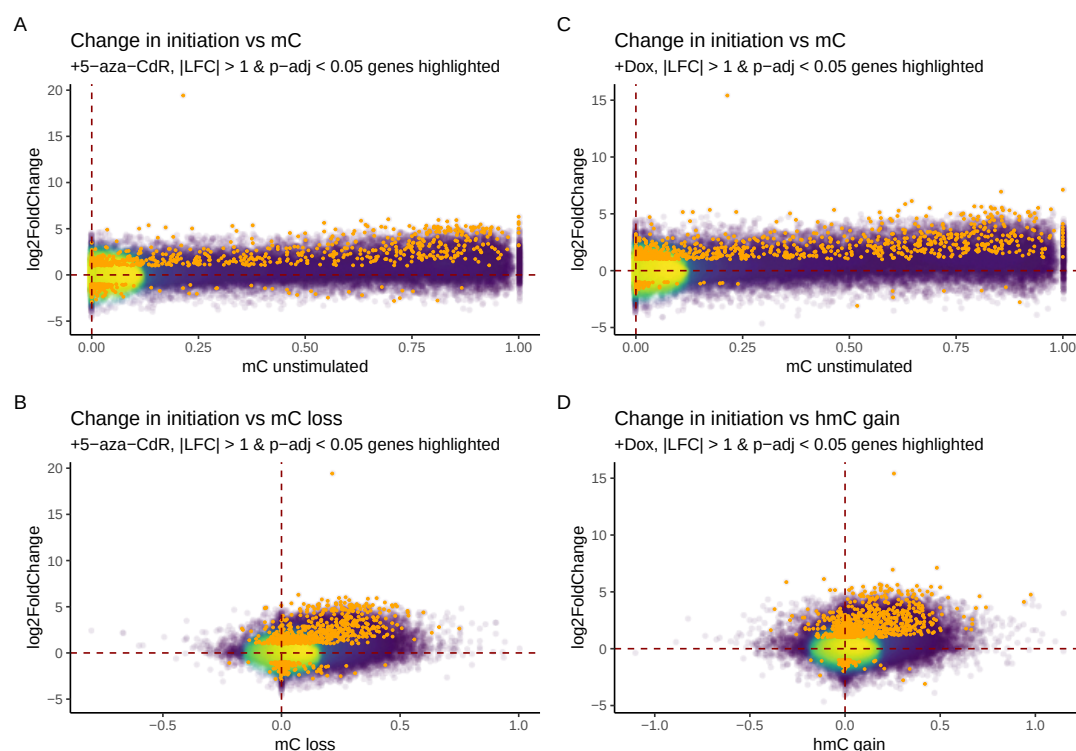


Figure 68: **Analysis of DNA modifications in initiation regions.** Density scatter plots comparing mean DNA modification level of change (log2FC) in initiation peaks. Significantly changed initiation regions are highlighted in each case. A,C. Mean mC in the unstimulated sample versus log2FoldChange after +5-aza-CdR and +Dox treatment. B. Mean mC loss versus log2FoldChange after +5-aza-CdR treatment. D. Mean hmC gain versus log2FoldChange after +Dox treatment.

Using the peak annotation defined previously (Section 4.2.10.2) we split the peaks into peak classes (any regions with read counts below 10 reads were filtered out). TSSs

⁷⁸For log2FC with p-adj < 0.05 they increase to 5.44 and 8.50, respectively.

and TSS-proximal regions mostly lost PRO-Cap signal upon +5-aza-CdR treatment, whereas there were more gains after +Dox (Figure 70A-C) suggesting that hmC at the TSS may be more permissive than demethylation. However since TSS-proximal regions are generally unmethylated (Figure 69A), these effects are likely to be secondary (due to distal activity for example).

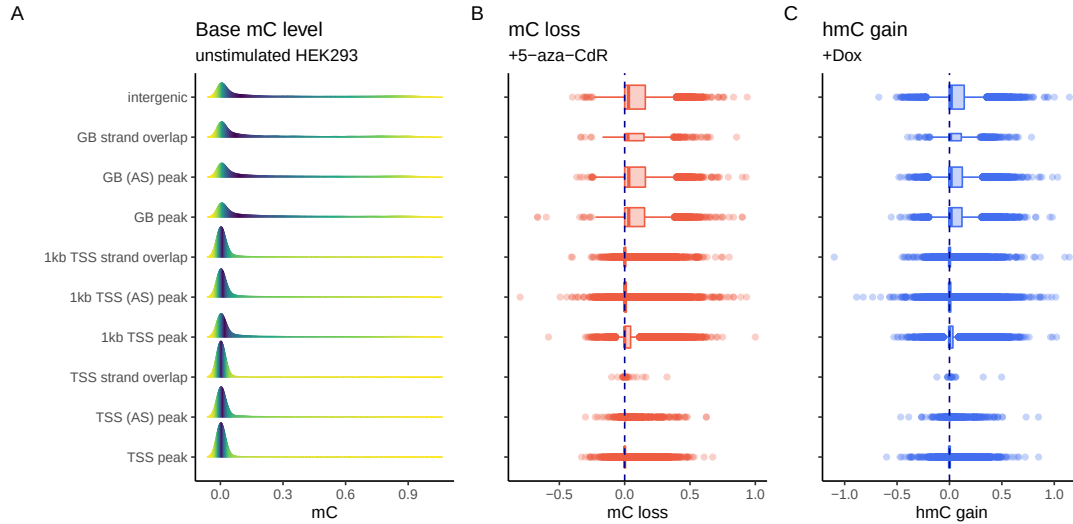


Figure 69: Analysis of DNA modifications in initiation regions. Density scatter plots comparing mean DNA modification level of change ($\log_2FC$) in initiation peaks. Significantly changed initiation regions are highlighted in each case. A,C. Mean mC in the unstimulated sample versus $\log_2FC$ after +5-aza-CdR and +Dox treatment. B. Mean mC loss versus $\log_2FC$ after +5-aza-CdR treatment. D. Mean hmC gain versus $\log_2FC$ after +Dox treatment.

Gene bodies are generally methylated (recall metagenes in Figure 30), with hmC levels related to expression [54, 144, 165, 120], thus it was interesting to see that initiation regions in gene bodies were dominated by upregulation after both treatments (Figure 70A-C). Gene body peaks lost the most mC and gained more hmC compared to TSS-proximal regions (Figures 69B and C) suggesting that loss of DNA modifications may promote spurious gene body initiation.

260 peaks were significantly changed (in the same direction) in both conditions (see examples in Figure 72), the directionality was consistent whenever a peak showed significant changes in at least one treatment (Figure 71A and B). However, there were

some examples of treatment-specific significant changes (see examples in Figure 73). Focussing on the peaks with minimal detectable differences from unstimulated in one condition but significant gains after the other, let us define a stable peak as one having an adjusted p-value exceeding 0.9, then we have 9 genes which are +5-aza-CdR specific⁷⁹ and 7 genes that are +Dox specific. In order of increasing p-adj: SOX3, TBX4, HSF2BP, TET1, CRT3-AS1, RIPK4, AL023882.1.

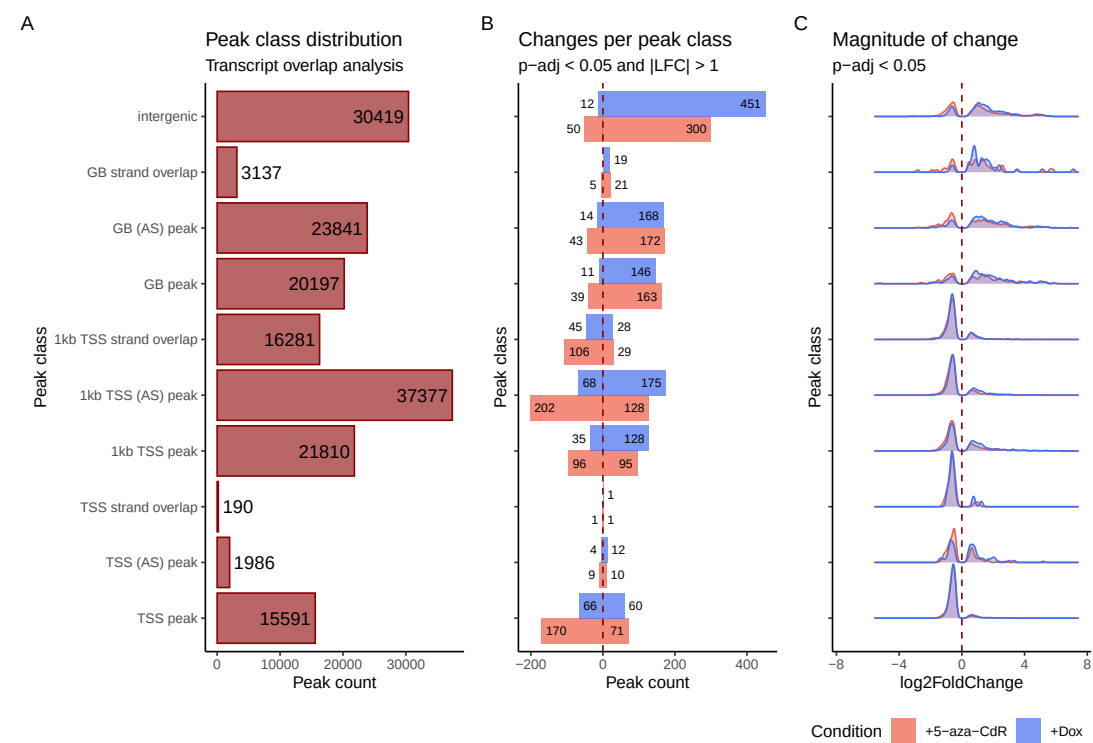


Figure 70: **Classifying differential signal.** A. Bar plots showing the number of peaks falling into each peak class in the DESeq2 results table. B. Bar plots counting the number of peaks with p-adj < 0.05 and |LFC| > 1 in each peak class upon +5-aza-CdR and +Dox treatment. Positive values count upregulated peaks, and negative values count downregulated. C. Density plots comparing the log2FoldChanges in p-adj < 0.05 peaks associated with each annotation-based peak class.

Only 17.8% of the peaks were intergenic, the rest overlapped with gene bodies, but intergenic peaks accounted for 38% and 30% of the upregulated peaks after +Dox and +5-aza-CdR treatment respectively (see examples in Figure 74). We wanted to see

⁷⁹In order of increasing p-adj: FBLN1, AC016831.6, LINC-PINT, EPHB2, ZNRFB3, KDM4B, CAPN5, EDN2, KCNK13, MFSD2A, CFHR5, RAB11A, AC084880.2

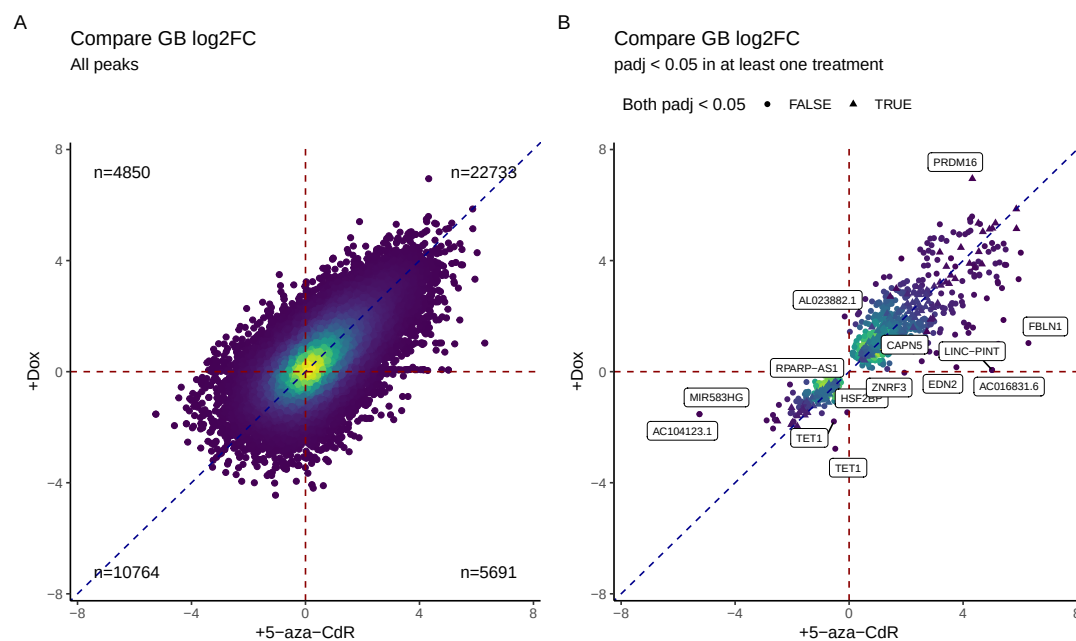


Figure 71: **GB initiation peaks comparison.** A. Density scatter plot comparing log2FC(+5-aza-CdR vs unstim.) and log2FC(+Dox vs unstim.), with each gene body (GB) initiation peak represented by a point. B. Subset of A, containing peaks with a significant change (p-adj < 0.05) after at least one treatment.

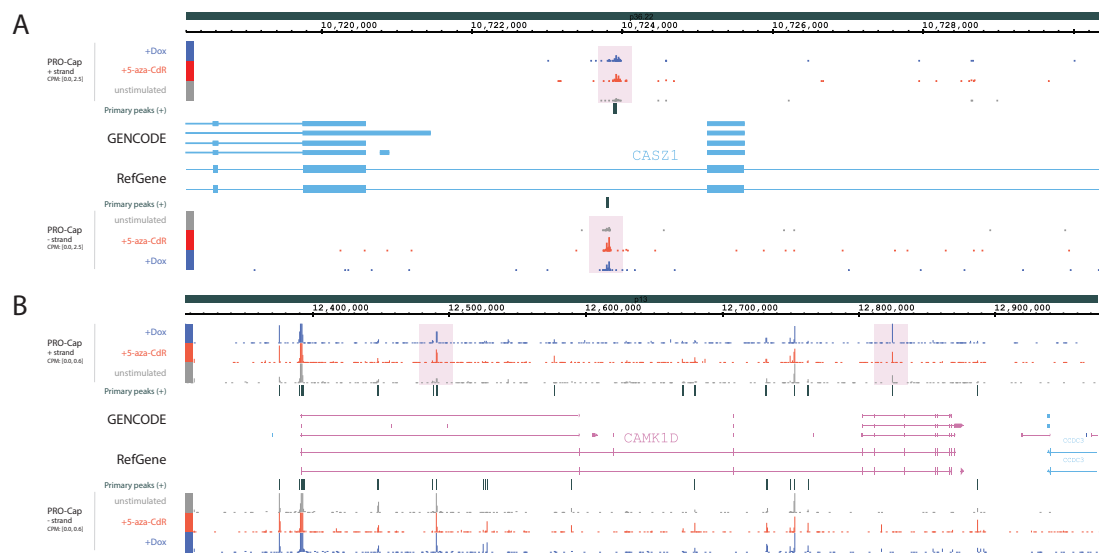


Figure 72: Examples of gene body peaks with changes in PRO-Cap signal. Genome browser tracks showing the combined (strand-specific) PRO-Cap signal, the primary peaks called on each strand, and a reference annotations from RefSeq and GENCODE. The orientation of the gene is shown as the colour, purple for plus strand and blue for minus strand and the peaks with changes in signal are highlighted in each case.

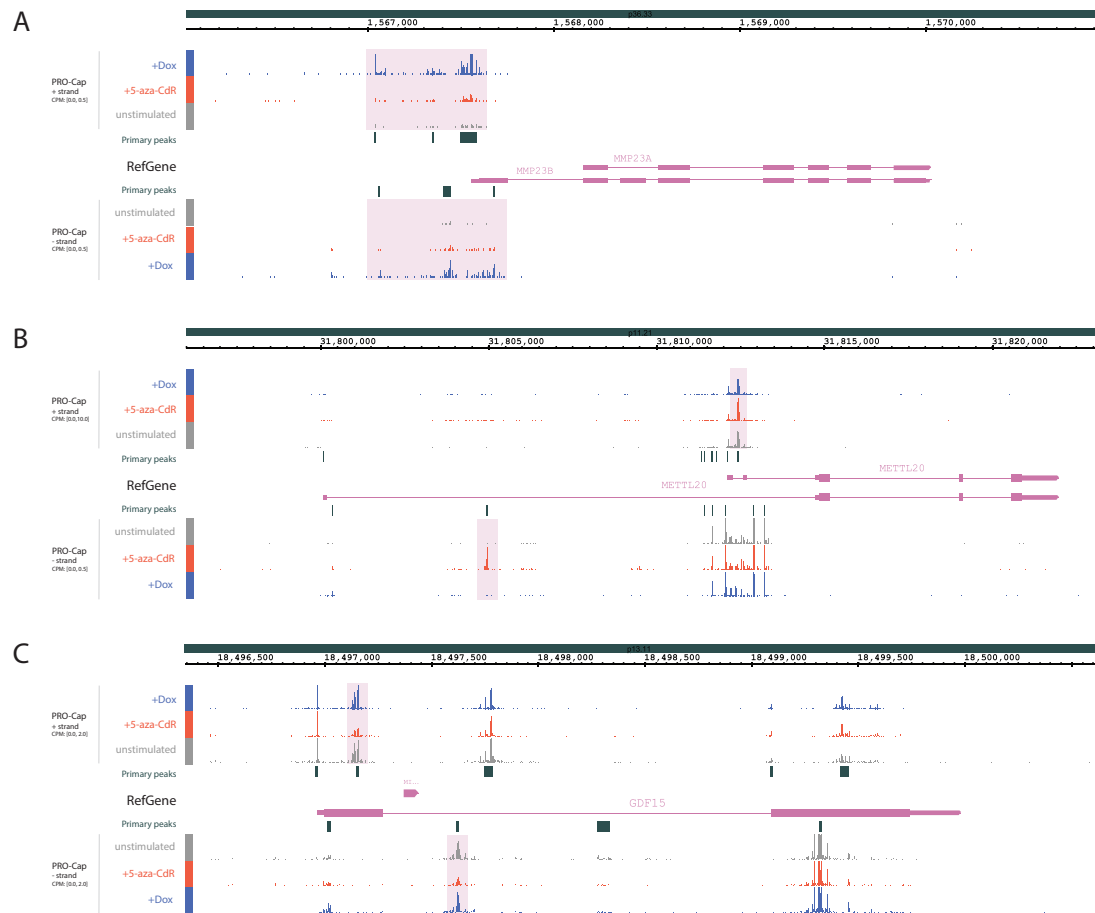


Figure 73: Examples of gene body peaks with treatment-specific changes in PRO-Cap signal. See Figure 72 caption.

whether these peaks were likely to be related to distal enhancers, and therefore checked for overlaps between public HEK293 ChIP data and the data-driven TSSs. Indeed, 11.7% of the intergenic peaks overlap with active enhancers (defined as those with both H3K4me1 and H2K27ac, Section 2.1.3) but only 0.5% of these show changes in PRO-Cap signal (Figures 75 A and B). We checked for gene associations for these enhancers using the EnhancerAtlas v2.0 [73], but didn't find any known associations.



Figure 74: Examples of intergenic peaks with changes in PRO-Cap signal. Genome browser tracks showing the combined (strand-specific) PRO-Cap signal, the primary peaks called on each strand, and a reference annotations from RefSeq and GENCODE. The orientation of the gene is shown as the colour, purple for plus strand and blue for minus strand and the peaks with changes in signal are highlighted in each case.

4.2.11.2 Analysis of termination, pausing, and uniformity

4.2.11.2.1 Termination

Unlike the TSS, the TES is much more difficult to define using PRO-Cap and PRO-Seq data, so instead we rely on the annotation as previous studies have. Some studies have

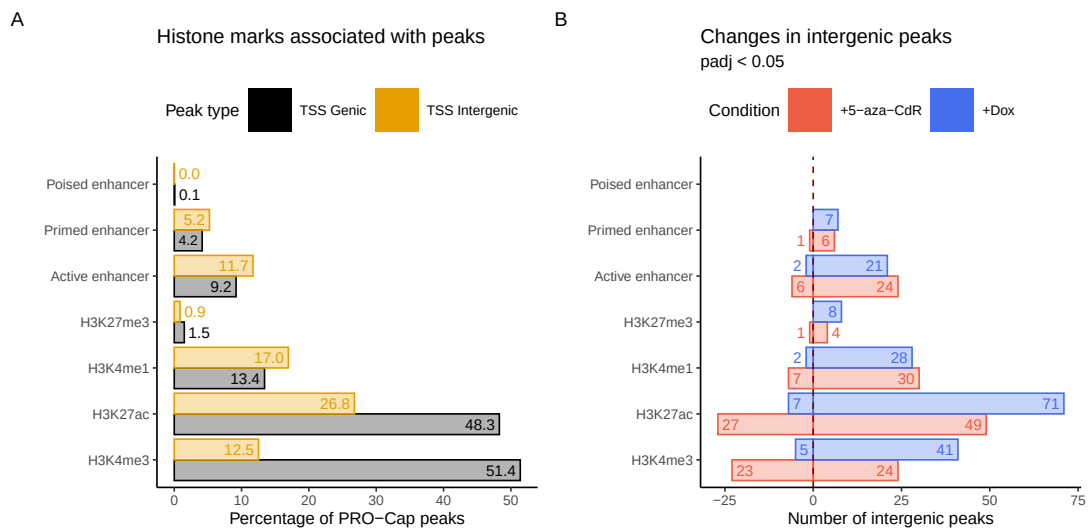


Figure 75: Intergenic peak analysis in the context of enhancers. A. Bar plot showing the percentage of genic and intergenic peaks that overlap with each category. We define active enhancers as regions with H3K4me1 and H3K27ac, primed enhancers have H3K4me1 signal but no H3K27ac or H3K27me3, and poised have H3K4me1 and H3K27me3 but no H3K27ac. B. Intergenic peaks with significant changes after treatment. Positive count for upregulated peaks, negative counts for downregulated.

simply analysed the signal downstream of the TES to check for read-through [261, 176] whereas others calculated a termination index by comparing the read counts upstream and downstream of the TES [11, 270, 180, 23, 24]. We chose to use the two metrics proposed by Parua et al. [180]:

- termination index (TI) = $\frac{n_{0-500down}}{n_{0-500up}}$, where $n_{0-500up}$ and $n_{0-500down}$ are the number of reads in the 500bp directly upstream and downstream of the TES, i.e. if there's a pause after the TES then $TI > 0$.
- termination elongation index (TEI) = $\frac{n_{250-750down}}{n_{0-500up}}$.

The mean TI and TEI in unstimulated were 1.91 and 2.58 respectively, meaning that we expect an almost 2-fold increase in signal after RNAP passes the TES which then increases further downstream. We grouped the replicates for each condition and calculated TI and TEI per gene. We found no clear gains or losses in TI or TEI after +5-aza-CdR or +Dox treatments (Figure 77A). There is no systematic bias of gain or loss in TI or TEI, suggesting that demethylation does not alter TES-proximal RNAP stalling or dissociation based on these metrics.

4.2.11.2.2 Pausing regions and pausing index (PI)

Previous studies have analysed pausing using different approaches with some being fully annotation based (comparing read count near the TSS with those in the rest of the gene body) [38, 11, 14, 176] while others are informed by the read density, e.g. bins with highest RNAP abundance [252, 52].

Since we took significant effort to classify regions of initiation, we chose to define the pausing index as

$$PI = \frac{\text{count}(m_i)/|m_i|}{\text{count}(e_i)/|e_i|} \quad (6)$$

where m_i is a major peak with length $|m_i|$ and e_i is the peak-independent part of the gene body, i.e. the elongation region less any 100bp-padded initiation peaks (both major and minor) with combined length $|e_i|$. By using this approach we hope that any

spurious gene-body initiation will not strongly bias the denominator counts⁸⁰.

Based on the PI metric used, the treated samples behaved similarly but showed systematically less pausing at the promoter compared to the unstimulated (Figures 77B-F). This is consistent with previous studies where mC is associated with increased pausing. This indicates either faster pause release (depletion in pausing but same elongation), lower initiation (fewer RNAPs recruited but same elongation), or more productive elongation (same initiation at the major peak but gain in elongation) in the treated samples.

Most TSS-proximal initiation regions show high magnitude losses after +5-aza-CdR (Figure 70C), thus faster pause release upon demethylation could contribute to the difference in PI. However, there was a preference for gains in TSS-proximal initiation after +Dox which is consistent with gains in productive elongation (possibly with contributions from gene-body peaks which consistently show more gains, Figure 70C). Therefore both explanations are possible.

We combined the PRO-Cap initiation data and with the PI to see whether losses in PRO-Cap initiation were indeed consistent with faster pause-release upon treatment, indeed, peaks that showed significant losses in PRO-Cap signal were associated with genes with faster pause release (Figures 76A and B).

Some peaks that gained PRO-Cap signal showed gains in PI. This can occur when there is an increase in initiation without consequent changes in elongation. Of these peaks most, 93% and 86%, showed no significant changes in elongation after +5-aza-CdR and +Dox treatment, respectively, meaning that for this subset of genes changes in DNA modifications resulted in increased recruitment.

4.2.11.2.3 Stalling and uniformity index (UI)

Here we use a slightly different approach to analyse pausing based on the gene body signal. We hypothesised that the uniformity of elongation signal may be affected by changes in DNA modifications, and thus wanted to define a uniformity index (UI) to check for extra “bumps” in the data (which would correspond to paused/stalled RNAP). We decided to use a measure based on entropy as proposed by Li et al. [137] in their study

⁸⁰Changes in PRO-Cap are usually reflected in changes in PRO-Seq but do not reflect changes in productive elongation which we wish to be sensitive to in this case.

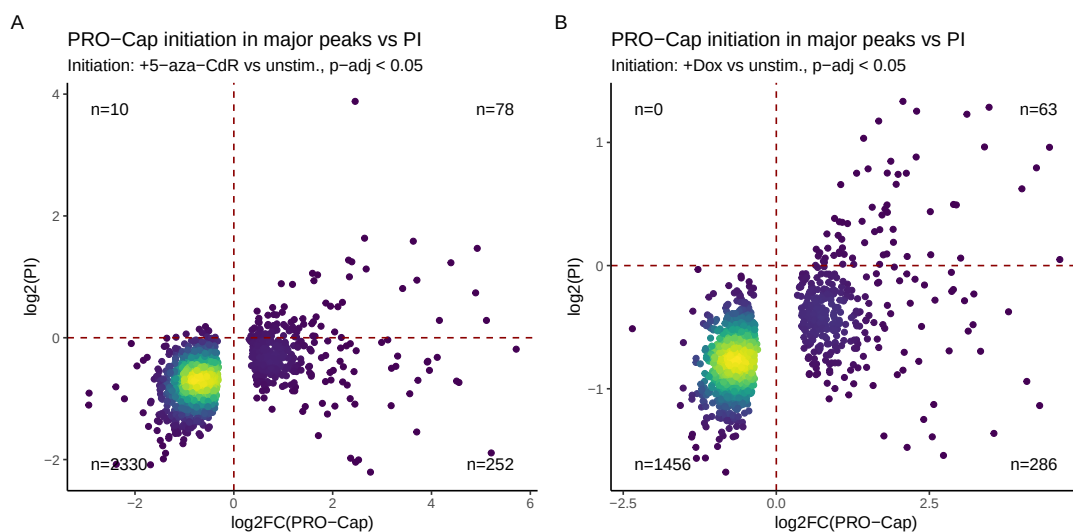


Figure 76: **Compare changes in PRO-Cap initiation in major peaks with change in pausing index (PI).** Density scatter plots comparing $\log_2\text{FC}$ in PRO-Cap signal in major peaks (those fulfilling the POM and POMCPB criteria) and gain in pausing index (PI) after treatment, i.e. $\log_2(\text{PI treated}/\text{PI unstimulated})$. The +5-aza-CdR vs unstimulated comparison is shown in A, and +Dox vs unstimulated in B.

of intron retention (Section 2.4.10). To calculate the UI we split the elongation regions into 20 equal bins and calculated the entropy (normalised by the maximum entropy of 20 values, 4.321928), thus genes with uniformity index ~ 1 will have a uniform/flat distribution whereas those with index close to 0 will show a lot of variation.

The summarised UI densities suggested slight differences between conditions: +5-aza-CdR results in the most smooth elongation, followed by +Dox, then unstimulated (Figures 77G and H). Therefore the presence of mC and/or hmC could indeed cause RNAP to stall slightly within the gene body which is consistent with the observation from the pausing analysis above (Section 4.2.11.2.2). We looked at the DNA modifications at base resolution to check for any relationships between methylation and UI and found that, indeed, most of the regions with low UI (the “bumpy” ones) were highly methylated (Figure 78A). The greatest gains in uniformity were associated with higher demethylation after +5-aza-CdR treatment (Figure 78B), highlighting a relationship between mC and stalling. The effect was less clear for mC loss after +Dox treatment and hmC gain and C gain (not shown) didn’t show a clear effect on stalling either

(Figure 78C and D).

4.2.11.3 Treatment-induced changes in elongation

We next compared the conditions to find if there were changes in engaged RNAP occupancy through the gene body upon genome-wide demethylation and/or hmC induction. Of the 14,281 elongation regions we noted significant ($p\text{-adj} < 0.05$) changes in 17% of the +5-aza-CdR treated and 5.4% of the +Dox treated, which suggests that demethylation is more disruptive to elongation. Similarly to the initiation analysis, we saw higher magnitude gains (Figure 79A and B). Again, the significant changes were consistent (gains or losses in both) but the number of candidate genes with $|LFC| > 1$ was significantly lower than we had in the initiation analysis with only 13 genes that were +Dox specific⁸¹ (Figure 79C and D).

We wanted to confirm that elongation regions which showed changes did indeed lose mC and/or gain hmC. By comparing the mean mC loss with elongation change after +5-aza-CdR treatment we found that genes with upregulated elongation did indeed lose more mC, and when grouping by the directionality of fold change there was a significant difference between the levels of mC loss (Figures 80A and C). We similarly compared the changes in elongation after +Dox treatment with mean hmC gain, and found a similar relationship whereby regions with high hmC gains were more likely to have upregulated elongation (Figure 80B and D).

PRO-Seq count is affected by both the speed and abundance RNAP which means that losses may be a consequence of either faster RNAP or lack of transcription (and gains can occur from either slower or more frequent elongation). Using the previous analyses, let us try to deconvolute these results:

- In the initiation analysis we noted gains in gene bodies in both conditions, which would fall into elongation regions and contribute to gains in elongation signal.
- Pausing analysis showed less pausing in both conditions, meaning that demethylation would increase the prevalence of pause release into productive elongation.

⁸¹VPS9D1-AS1, ALDH1A3, FGD5, **HSPA2**, TMEM130, C1orf167, NDUFB11, **LHX8**, MFSD4A, SERPINF1, **RFLNA**, **CALHM5**, and **TNFSF12**. Bold genes have $p\text{-adj} < 0.05$ in +5-aza-CdR, the rest don't reach the $|LFC| > 1$ threshold.

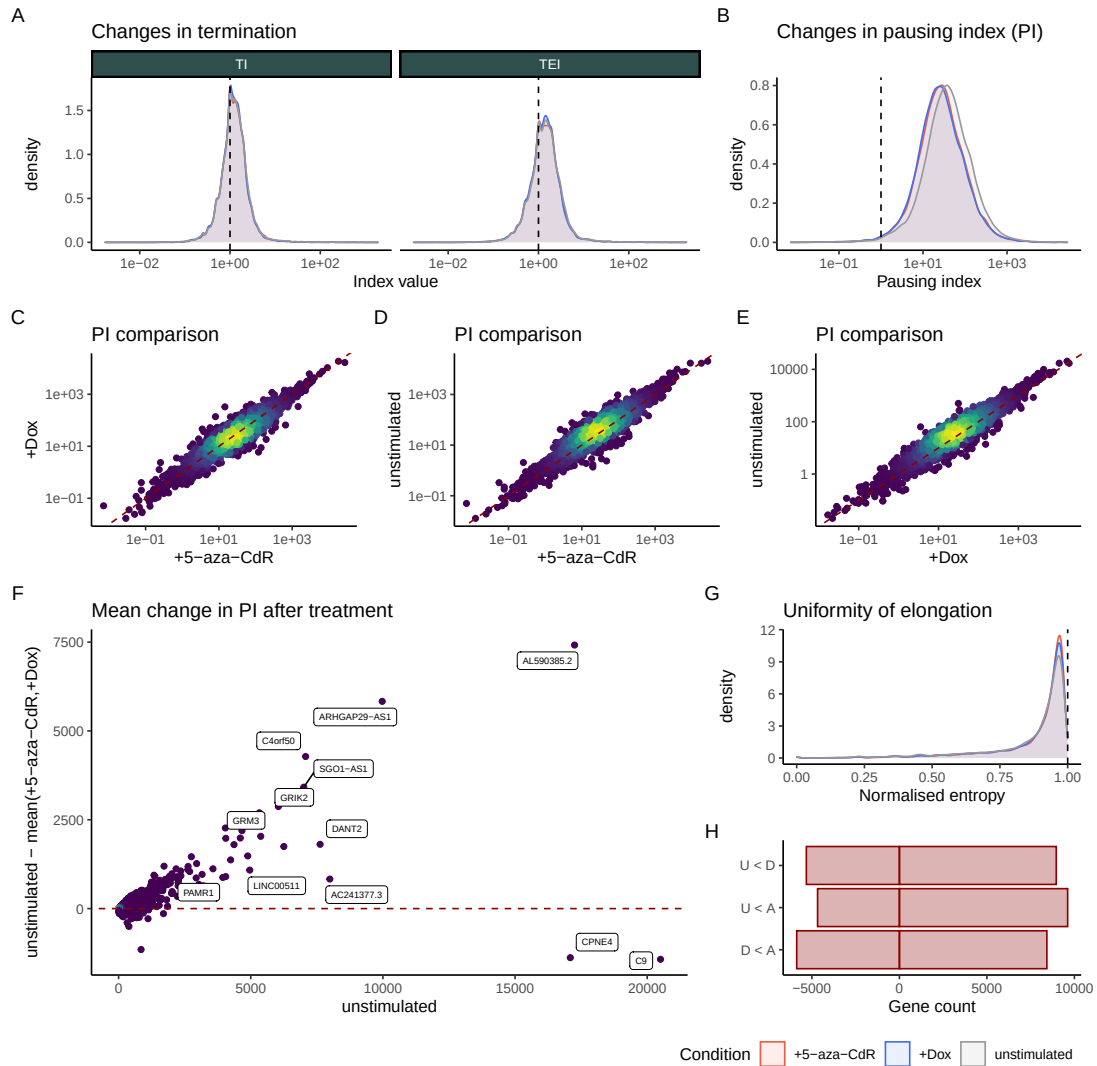


Figure 77: Termination, pausing, and uniformity of elongation. A. Density plots comparing the termination index (TI) and termination elongation index (TEI) between treatments. B. Density plot comparing the pausing index (PI) distribution for each treatment. C-E. Density scatter plots with each point representing a gene with defined elongation. The PI's for each treatment are compared directly, with a line of slope 1 highlighting the systematic differences. F. Scatter plot comparing the PI in unstimulated, with the mean change in the treated samples. G. Density plot comparing the uniformity index (UI) distribution for each treatment. H. Number of genes falling into each category (A = +5-aza-CdR, D = +Dox, U = unstimulated) with positive values representing the number of true cases, and negative values for false cases.

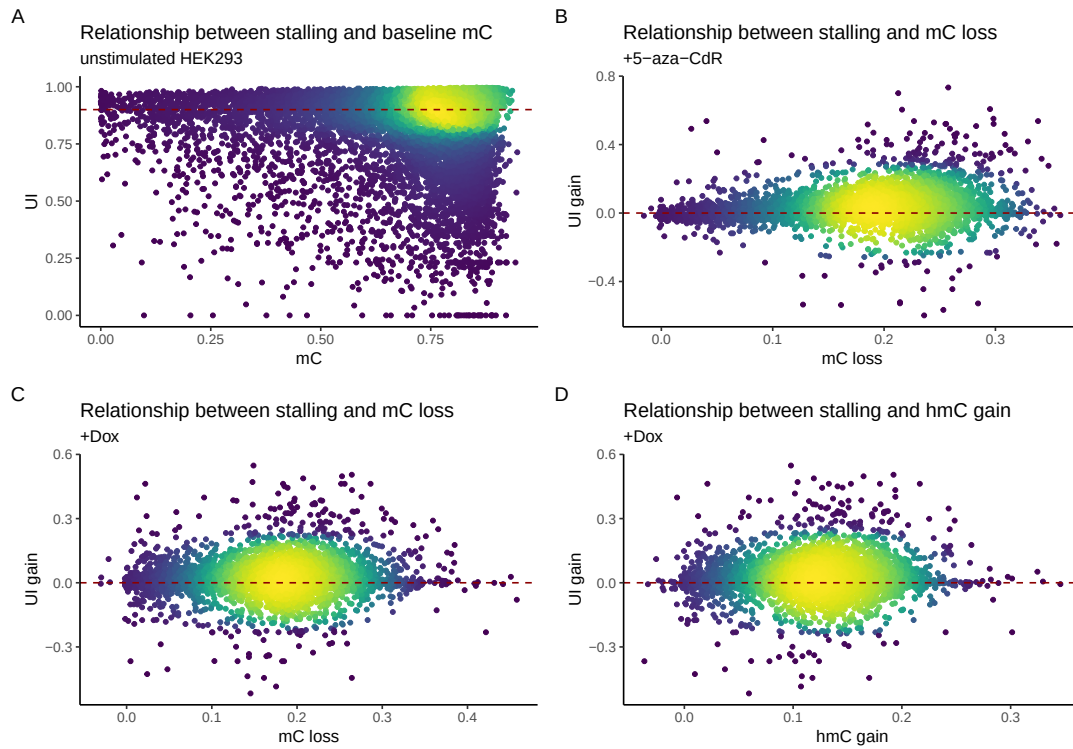


Figure 78: **DNA modifications and stalling.** Density scatter plots comparing mean DNA modification level or change and uniformity index (UI). A. The mC level (as measured by TAPS- β) in each elongation region versus the uniformity index (UI) in unstimulated. B-C. Mean mC loss after +5-aza-CdR treatment with the change in UI (+5-aza-CdR - unstimulated), and similarly for +Dox. D. Mean hmC gain after +Dox treatment compared with UI gain (+Dox - unstimulated).

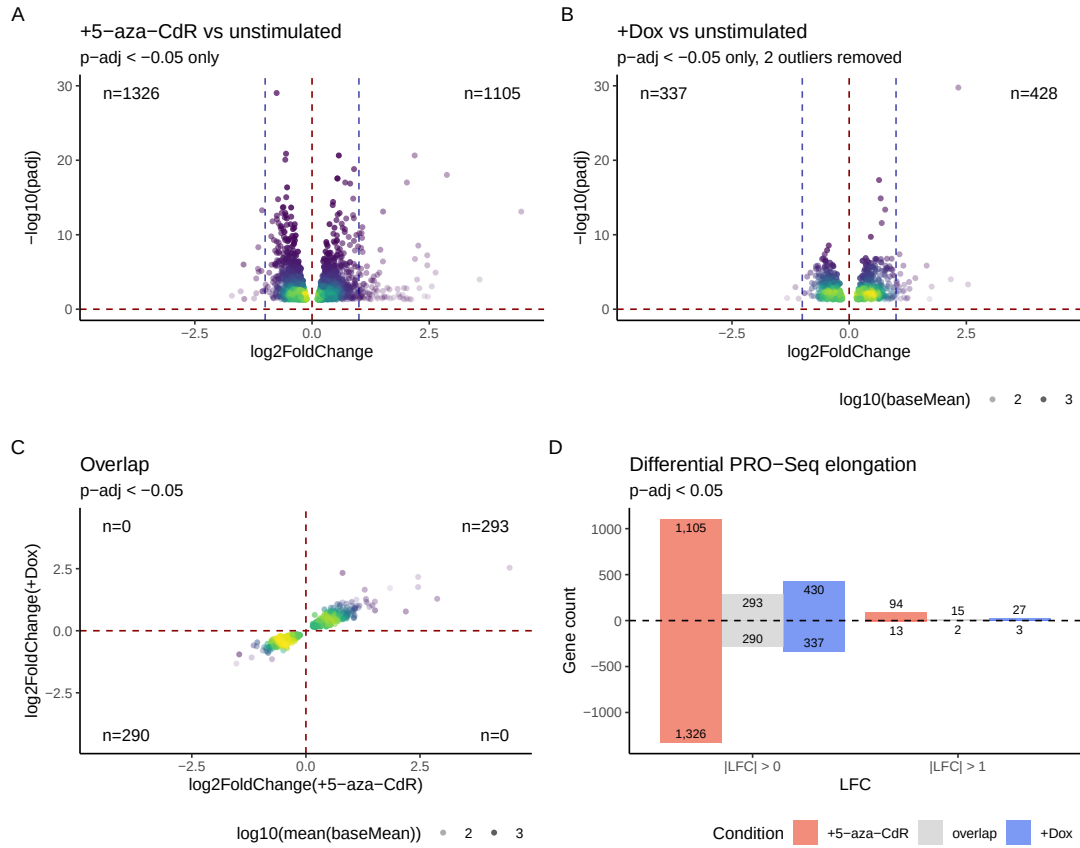


Figure 79: Differential signal analysis of all elongation regions. A-B. Volcano plots comparing +5-aza-CdR vs unstimulated and +Dox vs unstimulated, respectively. The density scatter plots show the log2FoldChange and -log10(adjusted p-value) for each peak with a significant (< 0.05) adjusted p-value, with some outliers removed through limiting the x- and y-axis for easier comparison. The counts in each quadrant are included, and dashed lines highlight $LFC = 1$ and $LFC = -1$. C. Density scatter plot comparing the log2FoldChange after each treatment for peaks that have p-adj < 0.5 in both. D. Bar plot counting the number of peaks that are upregulated (positive values) and downregulated (negative values) after each treatment subject to LFC filtering, with the overlapping peak set shown in gray.

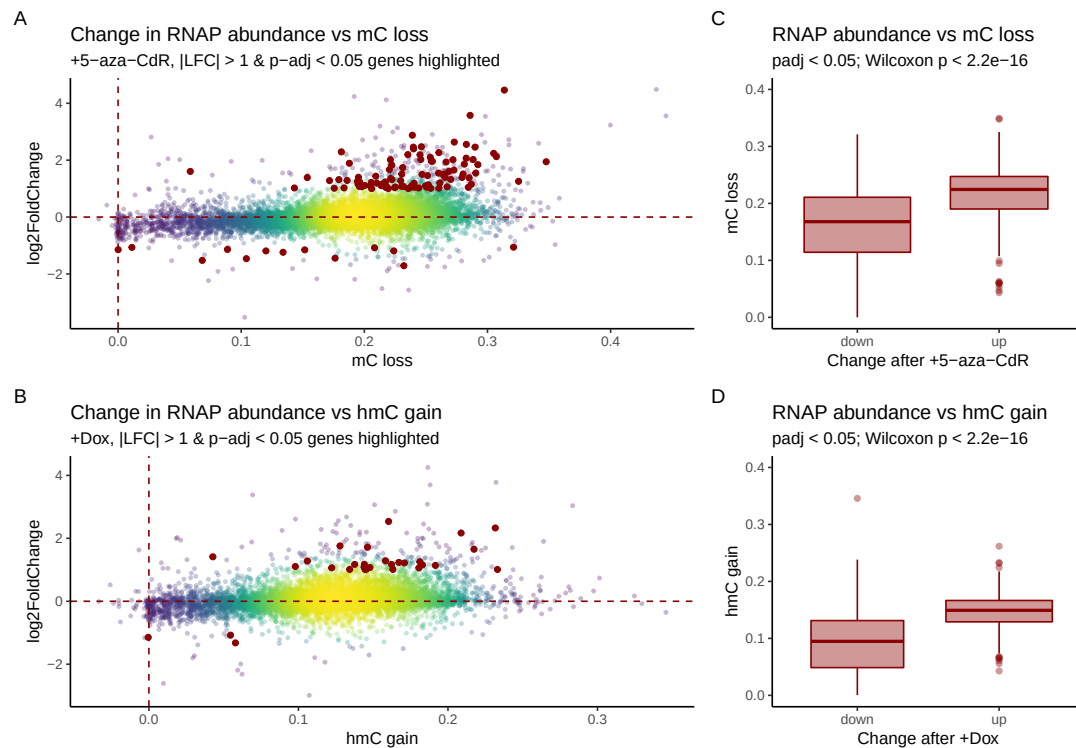


Figure 80: **DNA modification analysis of elongation regions.** A. Density scatter plot comparing mean mC loss in an elongation region with the log2FoldChange in PRO-Seq signal after +5-aza-CdR treatment. Elongation regions with significant ($p\text{-adj} < 0.05$ and $|LFC| > 1$) changes are highlighted. B. As in A, but comparing mean hmC gain with log2FoldChange after +Dox treatment. C-D. Boxplots grouped by the directionality of log2FoldChange of elongation regions with $p\text{-adj} < 0.05$, showing mC loss after +5-aza-CdR and hmC gain after +Dox treatments, respectively.

- The uniformity index analysis suggested most consistent (least bumpy) elongation in +5-aza-CdR, then +Dox, and finally unstimulated. This suggests that the gains in gene body initiation seen in PRO-Cap do not affect the total elongation count significantly.

In other words, the gains in elongation are more likely to be evidence of more productive (rather than slower) elongation and losses can be a consequence of either (speed or reduced transcription), but this can be de-convoluted using the RNA-Seq analysis.

In summary, elongation, as measured by the density of active RNAP occupancy downstream of major initiation regions, was affected by global changes in mC and hmC. Most of the high-magnitude ($|LFC| > 1$) changes were gains, with +5-aza-CdR treatment resulting in 109 upregulated elongation regions, whereas TET1-CD upregulation showed gains in 42 (15 of which were common to both treatments). Thus both mC loss and hmC gain were associated with increased elongation, but few genes showed strong changes. Through direct comparison with base-resolution mC and hmC measurements, we confirmed that the magnitude of mC loss and hmC gain in the elongation region was related to the magnitude of elongation gains. Combined with the previous analyses, we conclude that changes in DNA modifications can induce more productive elongation however further experiments with a temporal component should be considered to verify this (Section 4.3).

4.3 Discussion

In this chapter we wanted to quantify the effects of altered DNA modifications on active transcription. We defined data-driven TSSs using PRO-Cap data, and by combining this with the density of active RNAP genome-wide (measured by PRO-Seq), we were able to compare initiation, pausing, elongation, stalling, and termination between samples with mC and hmC changes.

Based on the spike-ins we found that the +5-aza-CdR and +Dox treatments did not induce changes in transcription genome-wide, i.e. neither treatment considerably increased (or decreased) transcription throughout the genome. A more focussed anno-

tated metagene analysis confirmed that the data was consistent with previous studies, but showed minimal differences in average RNAP abundance between control and the treated samples with the exception of TSS-adjacent signal, suggesting some differences in promoter-proximal pausing.

We chose to define data-driven TSSs based on the PRO-Cap signal rather than relying fully on genome annotation for multiple reasons. The most important aspect to consider was the goal of calculating elongation signal without including contamination from initiating RNAPs. We found that although some data-driven TSSs did indeed overlap with the annotation, many did not, and therefore could contaminate elongation signal and possibly bias the results. Separately, by identifying data-driven TSSs we could analyse whether DNA modifications affected spurious initiation within gene-bodies or intergenic regions (including enhancers).

We began with an annotation agnostic analysis (AAA). We found that the treated samples appeared to have more initiation peaks, these were less stable (lower concordance between replicates) and the higher proportion of low-abundance peaks suggested evidence of higher spurious transcription. To gain greater insight into where changes may be taking place, we classified peaks based on overlaps with GENCODE transcript annotation in a supervised analysis (SA). Upon classification, of $\sim 202\text{k}$ initiation peaks, most (163,830) were genic. However, only 15,528 could be classified based on direct overlap with annotated TSSs on the sense strand, an additional 26,355 were found within 1kb, and 25,239 were gene-body (but not TSS-proximal). We also noted significant numbers of peaks falling onto the antisense strand (2,059 TSS overlap, 41,235 within 1kb of TSS, 29,732 within gene body). 96.4% of peaks could be assigned to a single gene, thus only a small proportion were filtered out.

We compared the PRO-Cap initiation signal between conditions and noted the prevalence of low magnitude losses, mostly in TSS-proximal regions. These losses can be explained by faster pause-release in treated samples, which would result in more long reads that are likely to be missed in the PRO-Cap protocol. Another contributing technical factor may be RNAP recycling. In particular, given a finite abundance of RNAP, increases in initiation in one region must be compensated by decreases in RNAP in

other regions. Thus high-magnitude increases in initiation of some genes could cause a depletion of RNAP for the rest of the transcriptome. To check this, one could quantify the proportion of free RNAPs at the start and at the end of the 30 minutes in which biotins are incorporated by adapting the live-cell RNAP analysis used by Steurer et al. [228].

We also saw a tendency for high magnitude gains, mostly in gene bodies and intergenic regions. In an effort to classify some of this signal, we overlapped the peaks with publicly available histone modification data and found that 9.2% and 11.7% of the gene body and intergenic peaks overlapped with active enhancers, respectively. Since enhancers can affect recruitment and initiation directly, seeing changes in initiation without proximal DNA methylation losses could be a consequence of altered enhancer activity. In this project, we used public HEK293 datasets to annotate enhancer regions, but due to the limited number of these we don't have a measure of uncertainty for each region. The overlap between derived active enhancers and the HEK293-specific EnhancerAtlas enhancers ?? was only 2.6%, meaning that we couldn't take advantage of enhancer-gene mapping. Therefore, for a deeper understanding how transcription of enhancers and enhancer-gene interactions are affected by DNA modifications, it would be better to map the relevant chromatin marks in each sample separately and use a technique such as Hi-C [142] to help define causal relationships.

The lack of high magnitude promoter-proximal changes can be partially explained by the generally low levels of promoter-proximal methylation at active genes⁸² (recall the metagene plots in Figure 31A). The treatments are more likely to directly affect transcription if the positions have the potential to lose mC (if they're unmethylated in the unstimulated sample then +5-aza-CdR and +Dox will not alter the methylation level). This could be addressed in future experiments by using cell lines with evidence of hypermethylation, such as those with isocitrate dehydrogenase (IDH) mutations (which affect the active demethylation pathway) or CpG island methylator phenotype (CIMP).

Throughout this chapter we have mentioned some limitations of PRO-Cap, and although this technique can provide high resolution data, interpreting that data can be

⁸²Whereas gene bodies and intergenic regions are more likely to be methylated.

difficult. Our main reason for using PRO-Cap was to identify TSSs in these HEK293 cells. Due to biases against longer fragments during the protocol, the abundance of PRO-Cap reads at a TSS is not an absolute measure of recruitment and initiation, but often an under-estimate (see Section). Therefore, to analyse whether changes in DNA modifications affect recruitment and initiation more thoroughly (especially the first 20bp downstream of the TSS), an orthogonal technique such as RNAPII-ChIP would be useful to quantify recruitment more directly but with the limitations of no strand information and antibody specificity.

We used a combination of PRO-Cap and PRO-Seq data to study pausing genome wide as DNA methylation has previously been linked to pausing. Rather than relying on average estimates of pausing distance, we analysed PRO-Seq signal downstream of the AAA TSSs in the context of pausing distance and confirmed that most pausing occurs within 60bp of the TSS (and in some cases earlier). This made us confident that we would not contaminate our defined elongation regions with pausing signal. The AAA pausing analysis suggested some differences between conditions in terms of read-count downstream of TSS, namely that RNAP may pause earlier in unstimulated. This observation, combined with the low-magnitude losses in PRO-Cap suggested that changes in DNA modifications affected pause-release in our cells.

For a deeper understanding, we defined elongation regions using the classified peaks, and calculated pausing index (PI) using the PRO-Seq data. The treated samples had lower pausing index than unstimulated, but were very similar suggesting that the effects of +5-aza-CdR and TET1-CD induction are comparable in terms of promoting pause-release. A related analysis of stalling within gene bodies was consistent with this observation, as the unstimulated sample had the least uniform PRO-Seq coverage profiles on average. In this analysis +5-aza-CdR treatment had the most smooth profile, followed by +Dox, suggesting that hmC causes more stalling within the gene body than unmodified C. Direct comparisons of stalling and mC loss showed that the genes with highest gains in uniformity after +5-aza-CdR treatment did lose more mC, however the +Dox samples didn't show an equivalently high concordance. In contrast, we saw no changes in termination index or termination elongation index when using annotation to

define TES sites.

We used PRO-Seq signal to quantify elongation. Similarly to PRO-Cap, both the AAA and SA analyses showed mostly losses in PRO-Seq signal upon treatment, but most of the high magnitude changes were gains. In both cases +5-aza-CdR was more disruptive than TET1-CD induction with 238 (51) and 67 (9) upregulated (and down-regulated) AAA 10kb elongation regions, respectively. In the SA, 136 elongation regions were significantly upregulated (p-adjusted < 0.05 , LFC > 1), of which 15 were affected by both treatments, 94 were +5-aza-CdR specific, and 27 were +Dox specific⁸³. This is consistent with the analyses of Veloso et al. [244], who found that gene body mC density is negatively correlated with elongation rate.

In conclusion, the treatments did appear to induce some changes in actively transcribing RNAP occupancy, however these changes were modest. Although we see evidence for decreased pausing, few genes showed differences in elongation which is the best predictor of mRNA levels of the quantities we've analysed in this chapter. However, due to some limitations regarding gene overlaps we couldn't analyse all genes so it is possible that we have missed some candidates. In our analyses we chose to filter with a focus on specificity to reduce false positives from gene contamination, but this obviously results in fewer gene candidates to analyse.

⁸³Only 18 genes were significantly downregulated, 2 by both, 13 by +5-aza-CdR only, and 3 by +Dox only.

5 Relationship between engaged RNA polymerase occupancy and mRNA abundance after genome-wide DNA modification changes

5.1 Introduction

Gene expression differs between individual cells, with the resulting mRNA population guiding the proteome which determines the cell's function, its mitotic state, and any responses to internal and external stimuli. Therefore measuring changes in mRNA is integral to understanding which processes are affected by any induced perturbations. For a long time it has been known that DNA modifications are one of the regulators of gene expression, primarily through their silencing potential [254, 173]. We now hold a more nuanced view of the role of mC and hmC, where demethylation can be associated with both positive and negative regulation. In a study of cell differentiation, Wan et al. [247] described a model where demethylation of two classes of methylated regions (each with distinct motifs) results in distinct effects on gene expression through allowing transcriptional activators and repressors to bind. Similarly, demethylation-mediated induction of miRNAs can disrupt expression of target genes [34, 194, 178].

In this project we focussed on determining how gene methylation (as opposed to distal regulation) affects transcription. On average, gene bodies are more methylated than intergenic regions [147] but unlike promoter-proximal methylation, gene body mC and hmC are positively associated with expression [54, 144, 165, 120, 241]. The reason behind this is unclear and some hypothesised (but debated) functions include protection from spurious transcription initiation [172, 236, 108], and regulation of splicing [118, 163, 256, 218, 146, 15].

This chapter is split into three sections discussing: the effect of DNA modification changes on gene expression, how specific levels of mC loss and hmC gain affect gene expression, and the relationship between RNAP abundance and mRNA levels. In the first section we discuss additional experiments that were carried out to measure DNA modification-independent effects on gene expression, and evaluate and compare how gene expression changes after demethylation and hmC induction. To present a broader

view of the changes induced, we contrast the effects of treatment on two cell lines, HEK293 and U2OS. Using the available base-resolution modification data, in the second section we evaluate how mC and hmC are related to expression, and how the magnitude of demethylation is related to the subsequent changes in gene expression. Finally, we contrast the two measures of transcription, active RNAP abundance and mRNA, to see whether the elongation changes calculated in the previous section propagate to the mRNA.

5.2 Results: mRNA analysis

5.2.1 mRNA abundance estimation

We want to evaluate whether the modest changes we noted in various parts of transcription affect mRNA abundance. Although mRNA abundance does correlate with RNAP elongation, high abundances of RNAP are not directly predictive of high gene expression. The majority of reads mapped in RNA-Seq are mature, spliced transcripts which can be used as a measure of gene expression.

The RNA preparation protocol used for these experiments involved extraction with an RNeasy Mini kit (74104, QIAGEN). Then, rRNA depletion was performed to maximise the proportion of relevant RNA transcripts. Otherwise, due to the high abundance of rRNA, much of the sequenced reads would have been filtered out. Note that poly-A selection was not performed as part of processing, this resulted in some nascent transcripts being part of the final libraries as well as some intronic counts⁸⁴.

When looking at some differential expression candidates (which were based on total gene-body signal), we found examples of genes with high intron (relative to exon) counts (Figure 81). High intron read abundance can be a consequence of misalignment, intron retention, or nascent or pre-mRNA RNA (which we would like to exclude). Therefore we restricted our focus to exonic reads. This also ensured that many of the counts associated with overlapping genes are significantly less contaminated with the overlapping neighbour counts since the actual exons rarely overlap (even if the gene bodies do)⁸⁵.

⁸⁴The decision to skip this step was made prior to PRO data being available.

⁸⁵Based on the annotation used, we reduced the 1,379,820 exons in the data into 338,462 overlapping exon clusters and of these only 11,144 had exons from two or more unique genes.

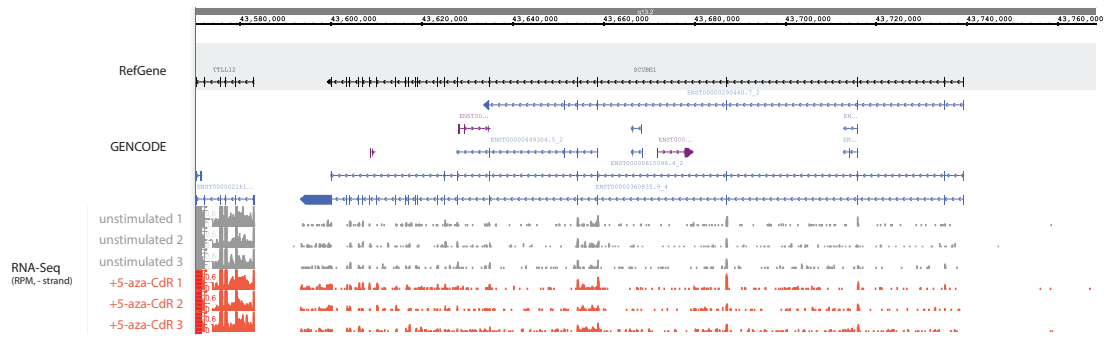


Figure 81: **High intron signal example.** Adapted genome browser view of RNA-Seq abundances falling into the SCUBE1 gene, normalised using RPM. Since the gene is on the negative strand, only the RNA-Seq counts on the negative strand are shown.

5.2.2 Experimental design : Mitigating DNA methylation independent treatment effects

Highlighting any changes in gene expression that were not consequences of altered mC and/or hmC levels was important since the two treatments worked in very different ways and had the potential to indirectly alter expression ⁸⁶.

HEK293 cells containing an inducible TET1 catalytic domain (TET1-CD) were the primary cell line used to investigate the effects of mC and hmC on mRNA abundance. However, to disentangle any hmC and/or demethylation independent effects that the TET1-CD may have, a second T-Rex HEK293 cell line containing a catalytically dead TET1-CD was included. Therefore, for the HEK293 differential expression analysis there are 4 groups of samples, i.e. TET1-CD +/- Dox and TET1-CDmut +/- Dox (Group 1 in Figure 82).

To establish the effect of demethylation, the cells were treated with 5-aza-CdR. In order to test whether the crosslinking associated with +5-aza-CdR could induce changes in gene expression (independent of any demethylating effect), the cells were incubated with 25 μ M 4-thio-2'-deoxythymidine (4-Thio-dT) and exposed to 365nm UVA light. This concentration resulted in comparable protein-DNA complex abundance to that found in the +5-aza-CdR samples **Pijus thesis**. Thus, in one set of experiments, we

Some of the multi-gene clusters are from families such as CCDC, HOX, KRT, and MIR.

⁸⁶5-aza-CdR is known to be quite toxic at high concentrations.

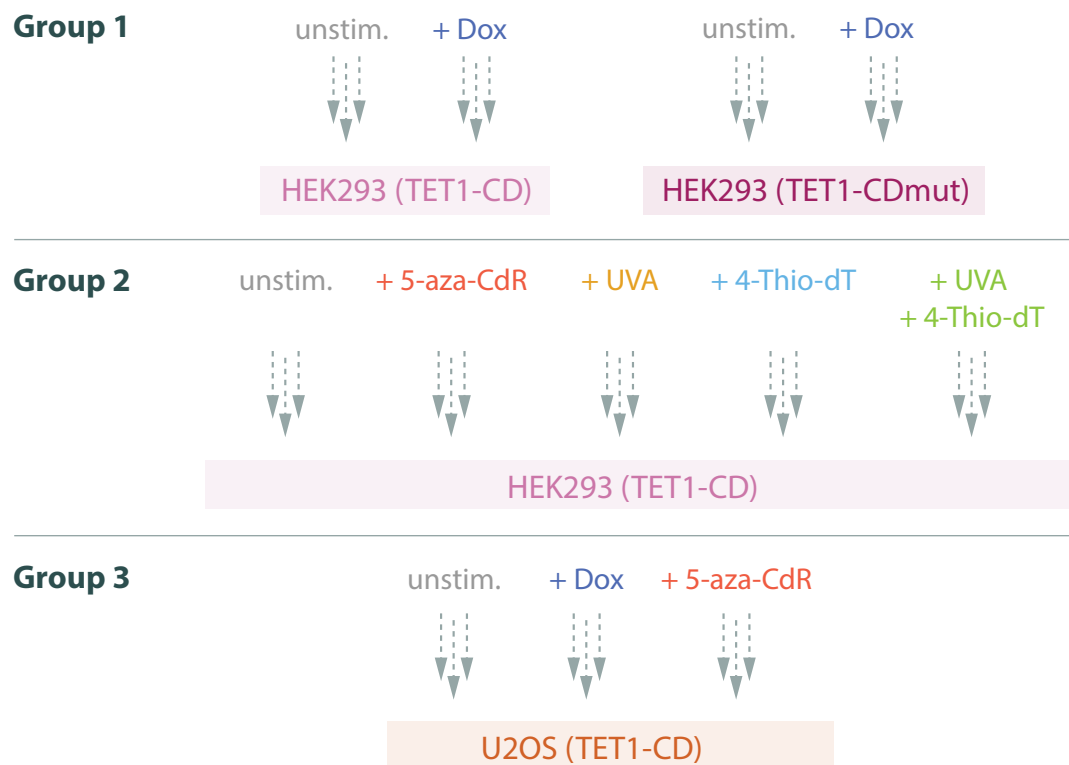


Figure 82: **Experimental set-up mRNA.** There were three sets of experiments carried out, with each treatment-cell type pair completed in triplicate. Group 1: Two cell types were included to compare the effect of catalytically inactive TET1-CD (TET1-CDmut) on expression. The TET1-CD or TET1-CDmut domains were induced with +Dox for 48 hours as before, with a set of untreated controls for each cell like. Group 2: To evaluate the possible effect of Aza-DNMT crosslinking, HEK293 TET1-CD cells were treated with 4-thio-2'-deoxythymidine (4-Thio-dT) and exposed to UVA (365 nm) such that the resulting DNA-protein crosslinks were comparable. Group 3: Another cell type, U2OS cells with a Dox-inducible TET1-CD, were treated with both +Dox and +5-aza-CdR to evaluate changes in a different cell line.

have HEK293 cells treated with 5-aza-CdR, 4-Thio-dT, UVA, UVA + 4-Thio-dT, as well as the untreated (Group 2 in Figure 82).

Lastly, in an effort to contrast HEK293 cells with another cell line, U2OS cells with a similarly inducible TET1-CD were used. As before, cells were treated with +5-aza-CdR and +Dox for 48 hours such that the levels of mC were comparable (Group 3 in Figure 82).

5.2.3 Count normalisation

Normalisation of mRNA counts is necessary to ensure inter-gene and inter-sample comparability. The software used, **salmon** [184], provides multiple options to help mitigate different sources of bias and variation but the choice should be informed by the data. Thus, to determine the effect of different normalisation strategies, we compared replicate count variability using PCA plots (Figure 83). Since the sequence normalisation (`--seqBias`) seemed to increase variability in some of the groups, we decided to stick with GC-normalisation as initially proposed in the DESeq2 vignette [153].

We also noted that the within-condition variability was very similar throughout, suggesting that using the full count matrix was sensible and would not inflate the expected estimates of dispersion significantly (which can happen through the inclusion of an unusually highly variable group, making the model less sensitive). In other words, to calculate the effect of +5-aza-CdR on HEK293 cells, the UVA, Thio., and UVA + Thio. samples were included when building the model before specifying the contrasts as simply the +5-aza-CdR and unstimulated. This was more nuanced in the +Dox analysis, as discussed in the next section.

5.2.4 Differential expression analysis

In the previous chapter we saw how DNA modifications affected different stages of transcription in HEK293 cells, here we wanted to see whether mRNA levels, estimated by RNA-Seq abundance, also showed changes in mC and hmC perturbed cells.

TET1-CD over-expression showed modest effects on gene expression, with 42 up- and 22 down-regulated genes ($\text{padj} \leq 0.05$) (Figures 84A and B). These changes

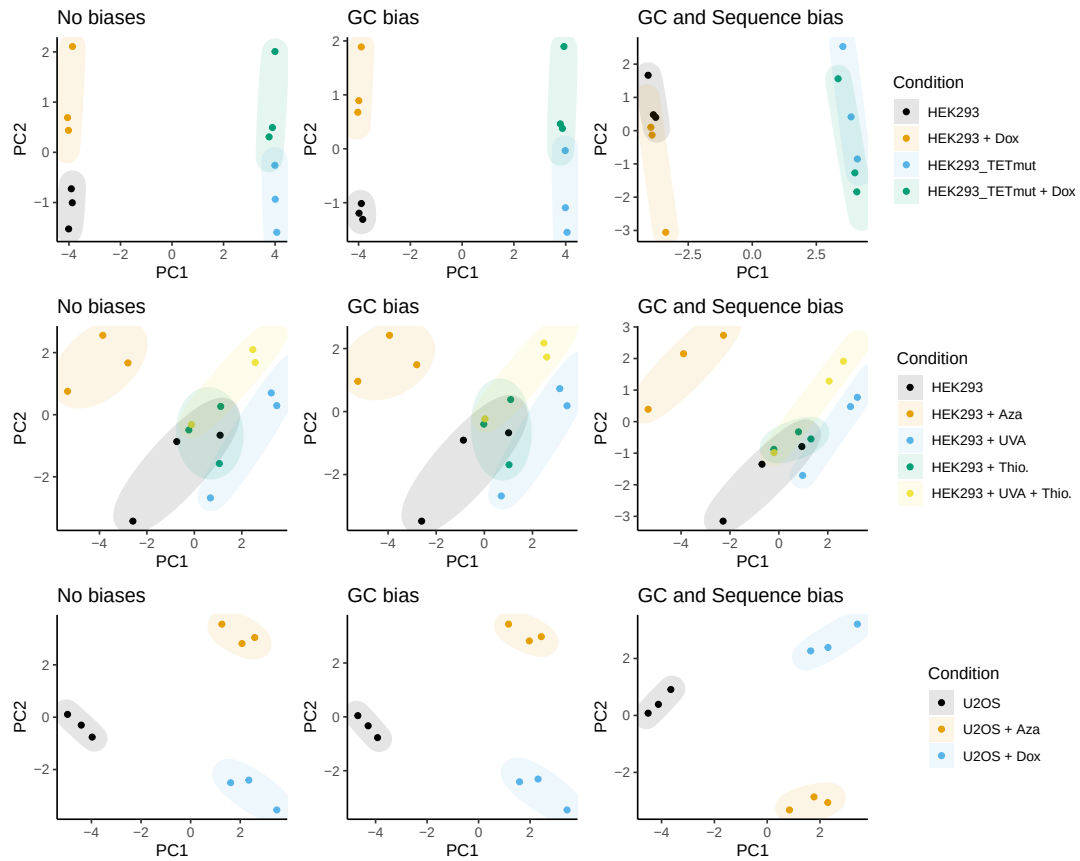


Figure 83: PCA plots of the normalised transcript counts comparing replicates after including GC bias, and GC and Sequence bias in the salmon model. The grouping of replicates is highlighted.

do appear to be a consequence of enzymatic activity, rather than off-target effects of +Dox induction based on the minimal overlap with mutant TET1-CD (mTET1-CD) deregulated genes. There were only 5 up- and 2 down-regulated genes when mTET1-CD was induced, with TET1 falling into the overlap.

Our goal is to determine the effect of TET1-CD induction, but the differences between expression in the mTET1-CD and TET1-CD samples meant that we could mask some of the smaller changes through inclusion of mTET1-CD in the count matrix since there were 927 genes upregulated in the unstimulated comparison, and 896 in the +Dox treated (Figures 84A and B). Indeed, some of the baseMeans⁸⁷ were inflated due to higher expression in mTET1-CD compared to TET1-CD (see Figure 85 for examples). Removal of the mTET1-CD samples from the count matrix resulted in 20 up- and 10 down-regulated genes, 23 of which overlapped with the original analysis and 7 new DEGs including ANKRD44, ANK3, EXT2, H1-4, GATA2, AGRN, and GATD3B with a mean adjusted p-value of 0.03 and mean absolute log2FC of 0.46. The new gene set did not highlight any new DEGs with highly significant changes so we kept the previous model with the full count matrix to increase sensitivity of detection at the majority of genes.

Treatment with +5-aza-CdR resulted in substantially more changes in gene expression, with 919 up- and 1082 down-regulated genes ($\text{padj} \leq 0.05$) (Figures 84C and D). Although we saw significant overlap between the DEGs (Figures 84C and D), accounting for 12% of the DEGs in the +5-aza-CdR vs. unstimulated comparison, the majority of changes in gene expression appear to be independent of the protein-DNA crosslinks and are more likely to be consequences of demethylation.

5.2.5 Comparing gene expression changes after +Dox and +5-aza-CdR

Recalling the PRO-Seq analysis in the previous chapter, we wanted to compare the magnitudes of gene expression changes after each treatment and determine how the two differentially expressed gene (DEG) sets compared.

Of the genes that showed changes after +Dox treatment, 65.6% were upregulated but

⁸⁷DESeq2's normalised measure of average abundance of a gene or region.

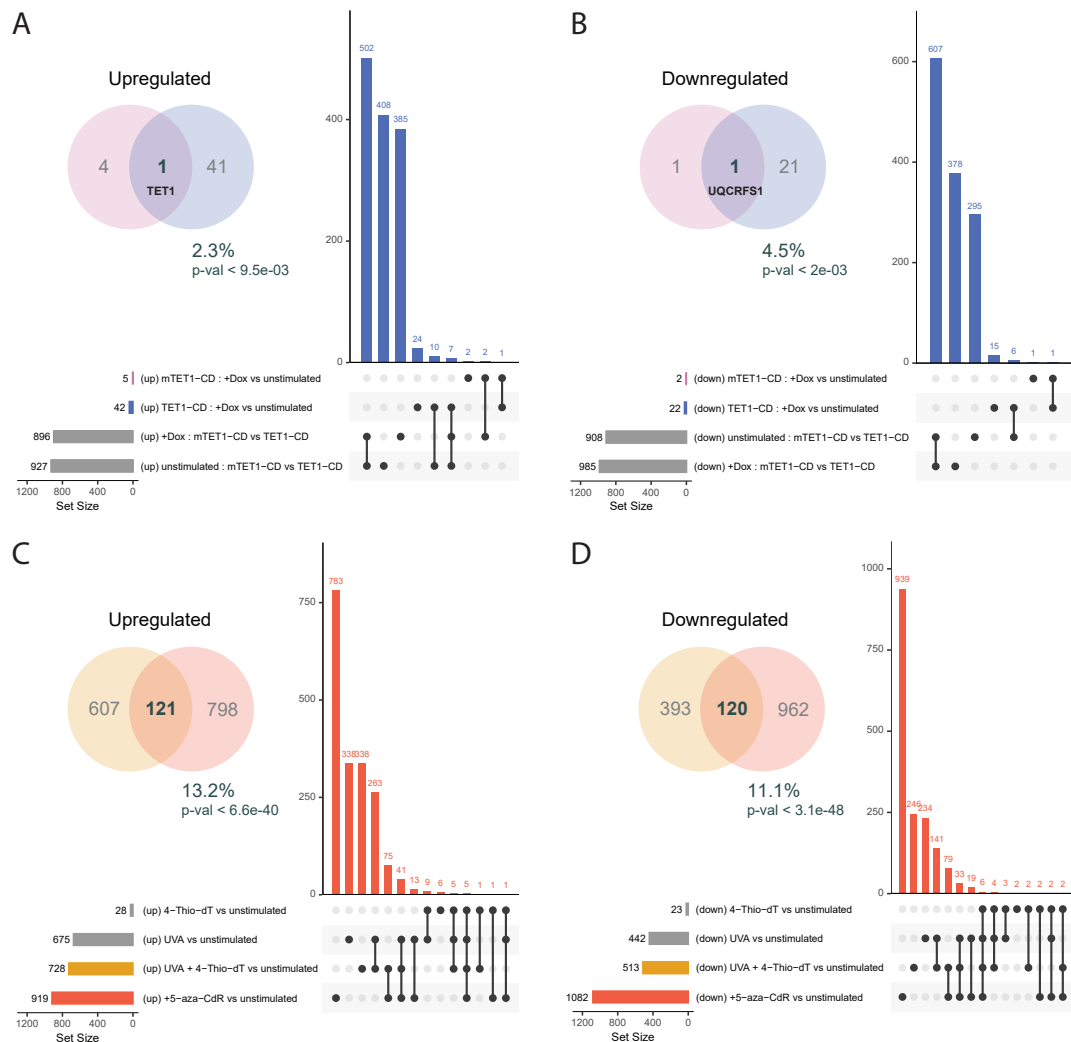


Figure 84: **Differential expression overlaps in HEK293 cells.** UpSet plots showing the number of overlapping genes between the DEGs in each comparison, defined as those with $\text{padj} \leq 0.05$ and $\log_2\text{FoldChange} > 0$ for up and < 0 for down, and corresponding venn diagrams showing the percentage and significance of overlap based on Fisher's Exact Test. Panels A and B show the effect of +Dox and mutant TET1-CD (mTET1-CD), and C and D compare the effect of +5-aza-CdR and crosslinking induced by UVA and 4-Thio-dT. The horizontal bars show the number of DEGs within each comparison, whereas the vertical bars count the number of DEGs in each overlap (the overlapping sets are shown as joined).

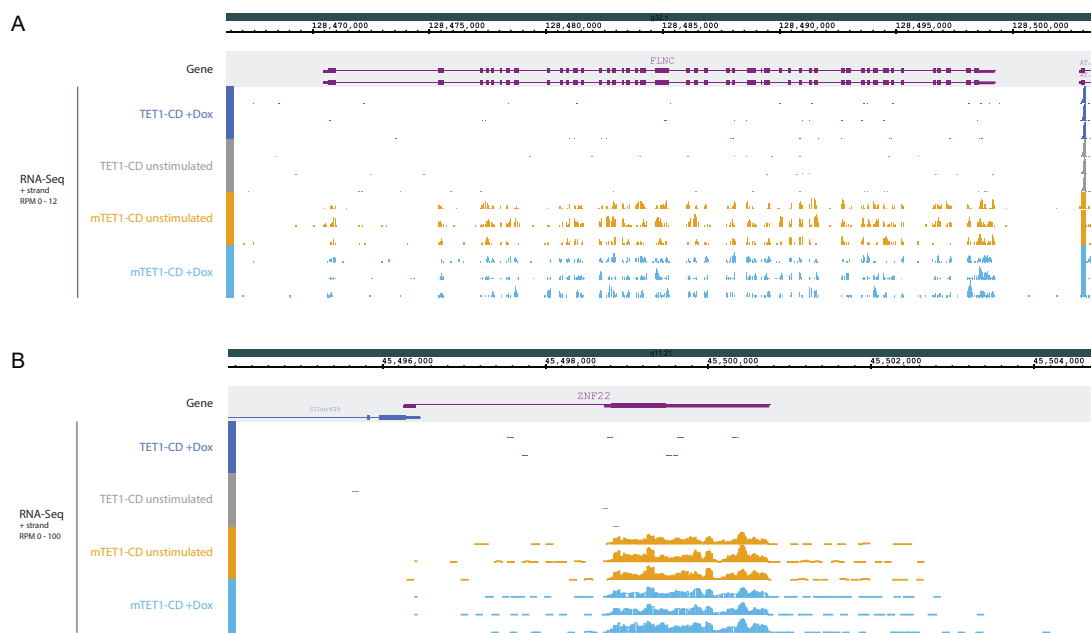


Figure 85: **Differentially expressed genes in HEK293.** A-B. RPM normalised RNA-Seq signal on the sense (+) strand in FLNC and ZNF22, respectively. The signal from each replicate is shown in its own track for each condition.

most didn't show high magnitude changes (the exception being TET1) (Figure 86A). In contrast, +5-aza-CdR treatment resulted in more downregulated genes, however most of the highest magnitude $\log_2FC$ changes were positive, with 248 up- and 19 down-regulated genes with $|\log_2FC| > 1$, consistent with a de-repression role (Figure 86B).

The overlaps between the DEGs were substantial, which was not surprising due to the partial demethylating effects of +Dox (Figure 86C). Of the 42 genes up-regulated after TET1-CD induction, 57% were also up-regulated after demethylation with +5-aza-CdR, leaving 18 whose expression gain could be explained by hmC induction only. The 22 down-regulated genes showed a smaller (but still highly significant) overlap, with 16 genes being +Dox specific.

We looked more closely at the DESeq2 results to ascertain whether the cut-off used was masking some associations. In Figure 87A we can see that most genes have similar average expression levels in the two sample groups (Group 1 and Group 2 from Figure 82), with outliers such as TET1 (which is expected) and ZNF22 (affected by mTET1

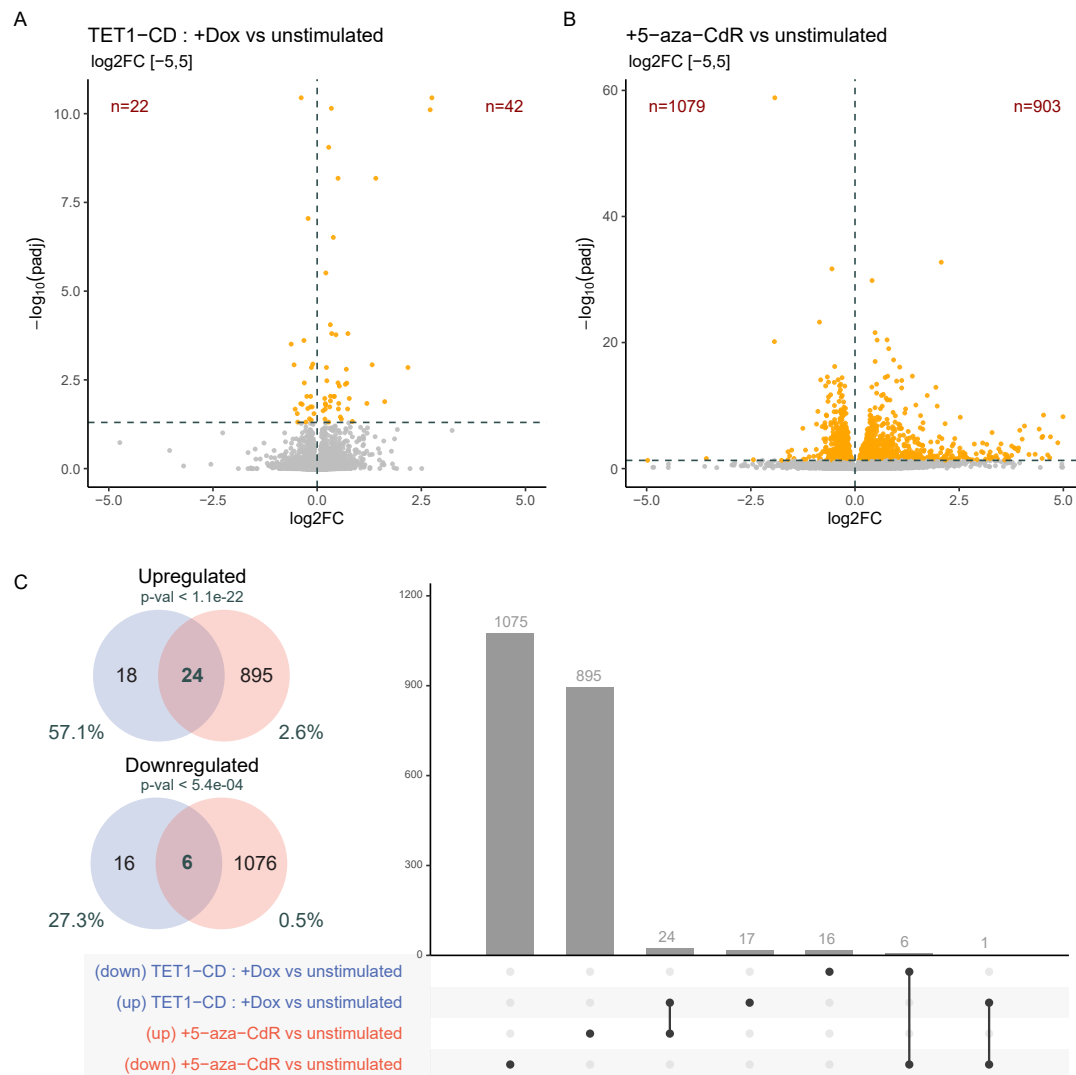


Figure 86: **Differentially expressed genes in HEK293.** A-B. Volcano plots which summarise the effect of +Dox and +5-aza-CdR on gene expression, respectively. In each case we count the number of significantly DE ($\text{padj} < 0.05$) genes in each direction. To improve visualisation, we ignore some gross outliers and limit the x-axis. C. We compare the DEGs directly using an UpSet plot, and summarise the results in venn diagrams.

expression as discussed previously).

Based on the little change seen after TET1-CD induction, it's unsurprising that the log2FCs show little correlation (Figure 87B). The +Dox-specific DEGs showed little correlation with log2FC after +5-aza-CdR treatment, suggesting that in most cases these genes are indeed affected by demethylation-independent TET1-CD activity.

Compared to HEK293 cells, U2OS cells showed more pronounced differences after +Dox and +5-aza-CdR treatment, suggesting higher sensitivity to changes in DNA modifications (Figures 88A and B). Again, +5-aza-CdR caused more deregulation compared to TET1-CD induction, and the overlaps were significant between the DEGs. The highest magnitude changes were positive after both treatments as before, therefore the general trends do not appear to be cell-line specific.

5.2.5.1 Focus on anti-sense lncRNA

Of the genes that were deregulated, we noted a pattern in the gene names which we wanted to explore further. In our dataset there are 16,751 unique lncRNA genes, of these 1,378 are “AS” (antisense) genes, which equates to ~8%. The differentially expressed lncRNAs seemed to be enriched for these anti-sense transcripts:

- HEK293
 - +Dox: 6 DE lncRNAs, 3 are AS (~50%)
 - +5-aza-CdR: 76 DE lncRNAs, 21 are AS (~30%)
- U2OS
 - +Dox: 36 DE lncRNAs, 9 are AS (~25%)
 - +5-aza-CdR: 89 DE lncRNAs, 17 are AS (~19%)

The number of DE AS transcripts is higher than expected if we were sampling lncRNAs randomly, and in some cases their corresponding sense genes are also DE (see example in Figure 89) but not always. Antisense transcripts have been reported as having a role in gene regulation through interference, so we compared the sense and antisense expression changes for genes with an annotated AS transcript (Figure

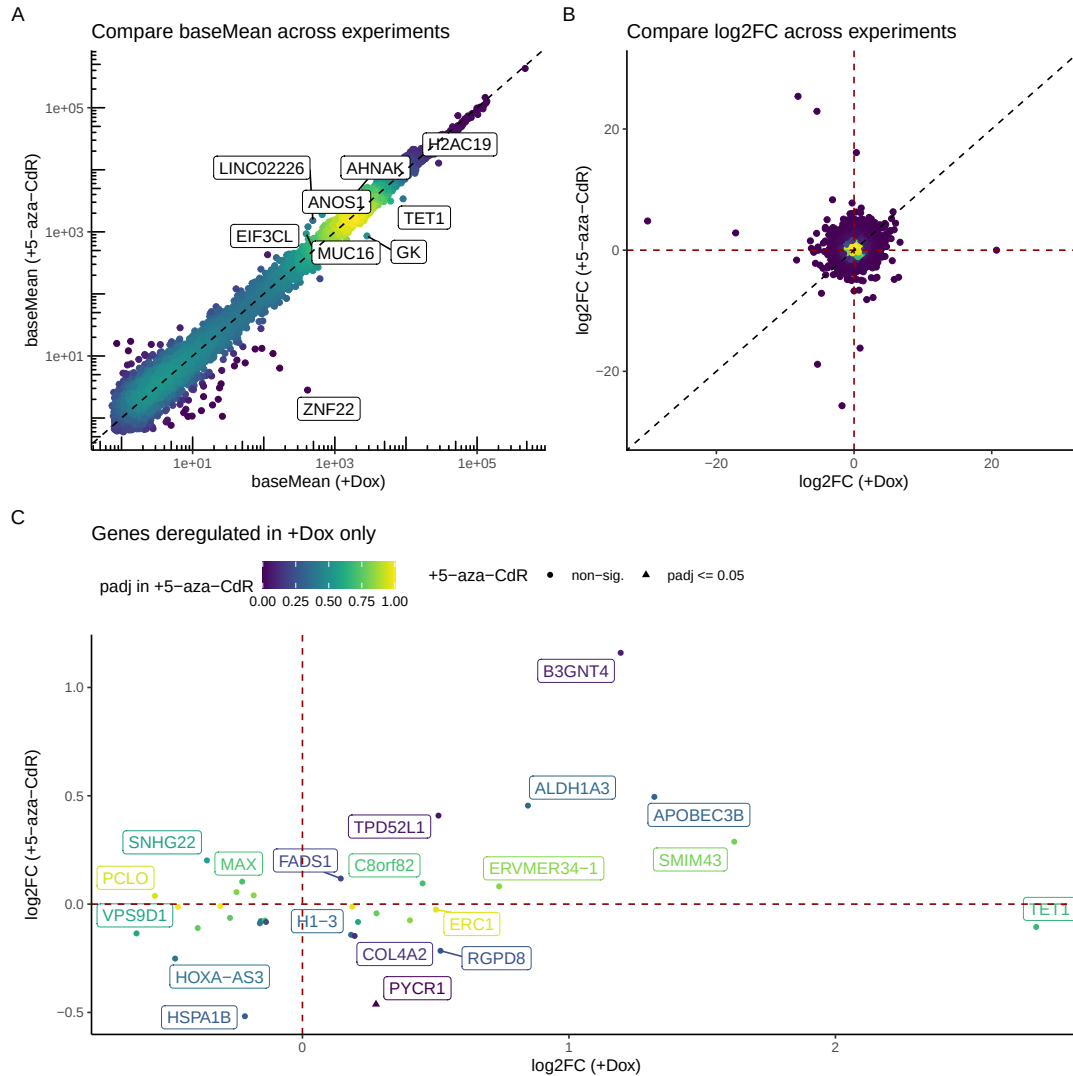


Figure 87: **DESeq2 output comparison of +5-aza-CdR and +Dox treatment in HEK293.** A-B. Volcano plots which summarise the effect of +Dox and +5-aza-CdR on gene expression, respectively. In each case we count the number of significantly DE (padj < 0.05) genes in each direction. To improve visualisation, we ignore some gross outliers and limit the x-axis. C. We compare the DEGs directly using an UpSet plot, and summarise the results in venn diagrams.

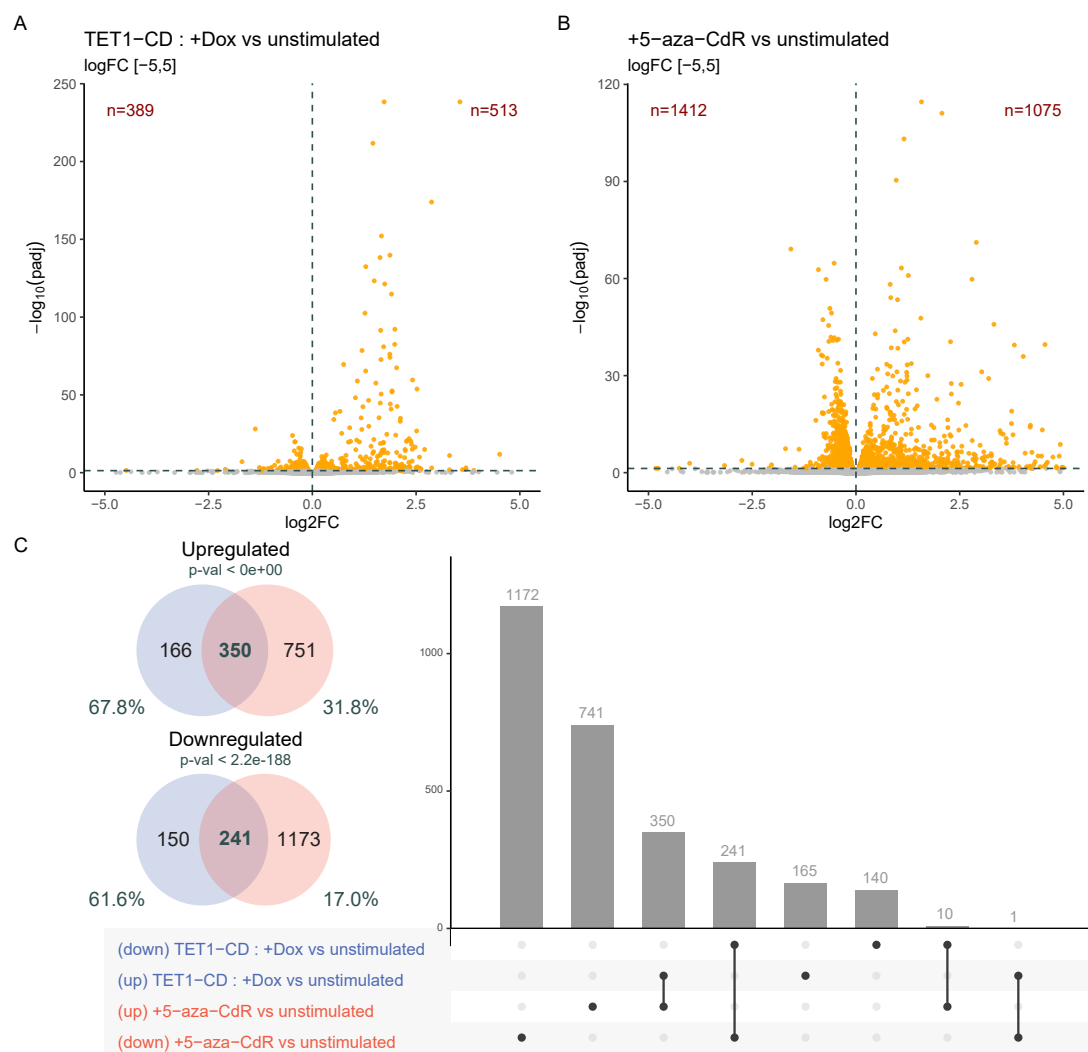


Figure 88: **Differentially expressed genes in U2OS.** See Figure 86 caption for details.

90). Based on this analysis there does not appear to be a strong relationship between directionality of change in our data, i.e. we do not see clear evidence of a silencing effect of AS transcripts.

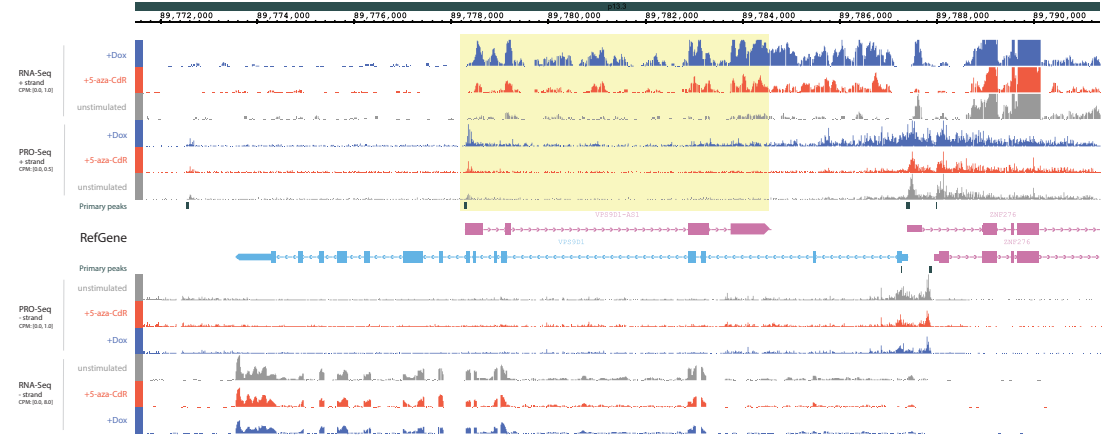


Figure 89: **Example of sense and antisense (AS) differential expression.** Genome browser view of VPS9D1 and VPS9D1-AS with tracks showing strand-specific PRO-Seq (RNAP occupancy) and RNA-Seq abundances (CPM normalised in each case). The predicted TSSs are also shown as Primary peaks. Note that the scales within a group are consistent, but differ between strands and experiments.

5.3 Results: Combining mRNA abundance and DNA modification data

In the previous analyses we noted that both +5-aza-CdR and +Dox treatment resulted in some expression changes. Here, we would like to analyse these changes in the context of gene methylation levels. In other words, how predictive are changes in gene body mC and hmC of mRNA abundance.

5.3.1 Relationship between baseline expression and mC and hmC levels

We first wanted to look at unstimulated HEK293 cells, to evaluate the relationship between DNA modifications and expression in unperturbed cells. Many groups have discussed the relationship between gene expression and mC and hmC density at TSSs and gene bodies (GBs). We can confirm that, in the HEK293 cells used in our ex-

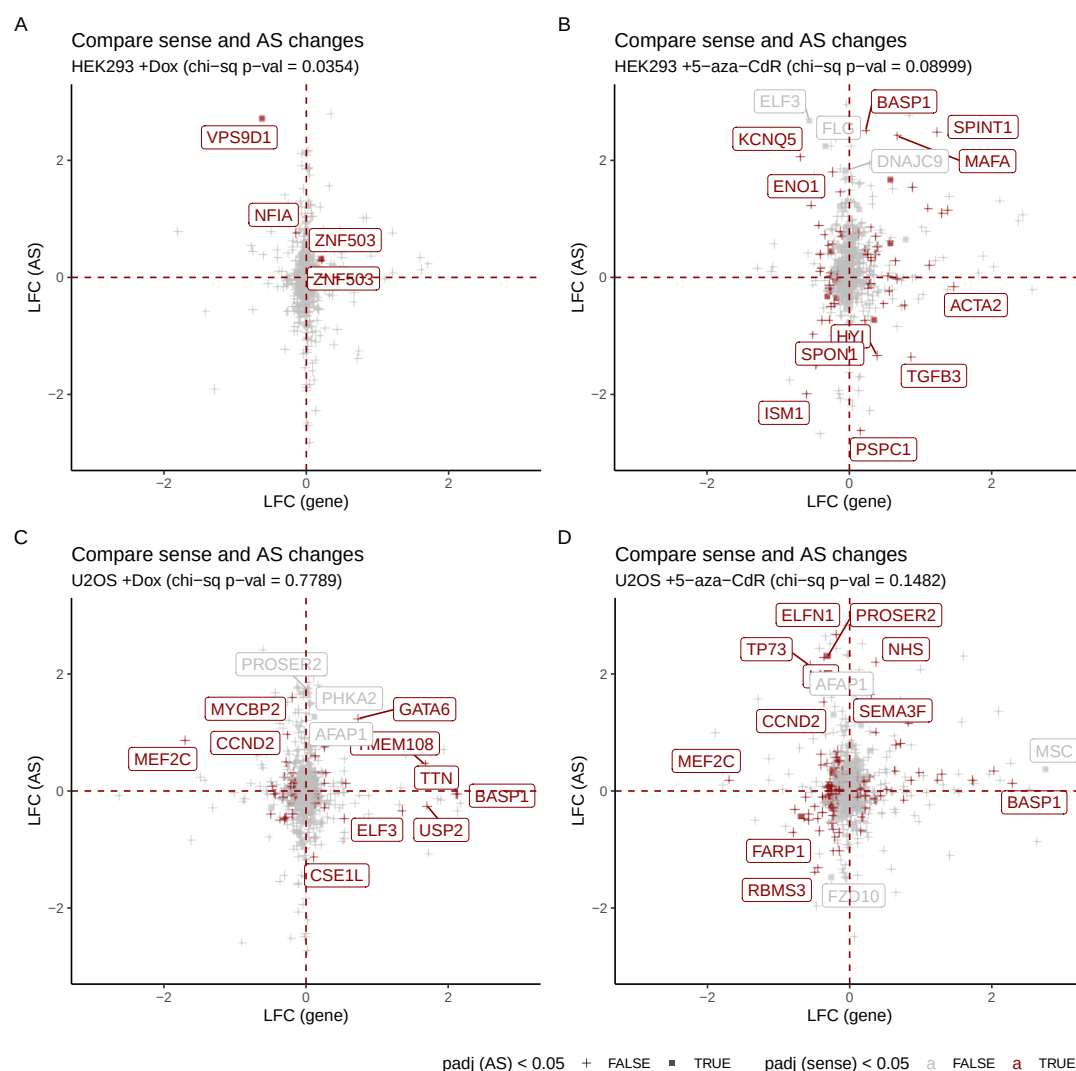


Figure 90: Sense vs antisense (AS) expression changes. Scatter plots with each point representing a sense - antisense transcript pair, e.g NFIA shows log2FC in expression of the NFIA gene on the x axis, and of the NFIA-AS gene on the y-axis. The shape of the points is determined by the significance of the AS transcript change, and the colour by the significance of the sense transcript change, i.e. a red square represents a gene with significant changes in both sense and antisense transcripts. Significance is calculated based on a chi-squared test of the quadrant counts. A-B. Changes after +Dox and +5-aza-CdR treatment in HEK293 cells. C-D. Changes after +Dox and +5-aza-CdR treatment in U2OS cells.

periments, the relationship between highly expressed genes and TSS hypomethylation followed by GB hypermethylation is evident (Figures 91A and B). Although the hmC density is more sporadic due to low baseline hmC content, there is still a clear split between the expression groups.

We then compared the mC densities after each treatment and found that the demethylation after +Dox and +5-aza-CdR had similar effects on all gene groups (Figures 91C and E). The hmC pattern in unstimulated and +5-aza-CdR cells was even more clear in the +Dox treated group (Figures 91B, D, and F). Since hmC density depends on initial mC level, highly methylated gene bodies (such as those in the high expression group) could gain more hmC than lowly methylated regions, resulting in a hmC density directly related to mC density.

5.3.2 Do genes with changes in expression show parallel changes in mC and hmC abundance?

We showed in Section 5.2.4 that the treatments did indeed cause changes in gene expression for a subset of genes, and above we have confirmed that there is a relationship between methylation and gene expression genome-wide. Now we would like to ascertain whether the differentially expressed genes (DEGs) showed changes in DNA modifications after +5-aza-CdR and +Dox treatment to see whether the *magnitude* of mC and/or hmC gain is predictive of the gene expression change.

To this end, we calculated the mC loss, hmC gain and unmodified C (uC) gain in each log2FC quantile across genes with p-value < 0.05 ⁸⁸ (Figures 93 and 94). The profiles in up- and down-regulated genes both show lower mC loss and hmC gain in gene bodies of the most altered genes, this is consistent with the observation that most highly altered (base on log2FC) genes are not highly expressed (Figure 92) and thus (based on the previous analysis) have lower baseline methylation.

The genes with strongest changes are also those that lose the most mC at the promoter (Figure 93), which makes sense as these were likely to have hypermethylated promoters to begin with and therefore were most susceptible to changes in expression

⁸⁸We used p-value rather than adjusted p-value to increase gene count in each group when looking at the general trends.

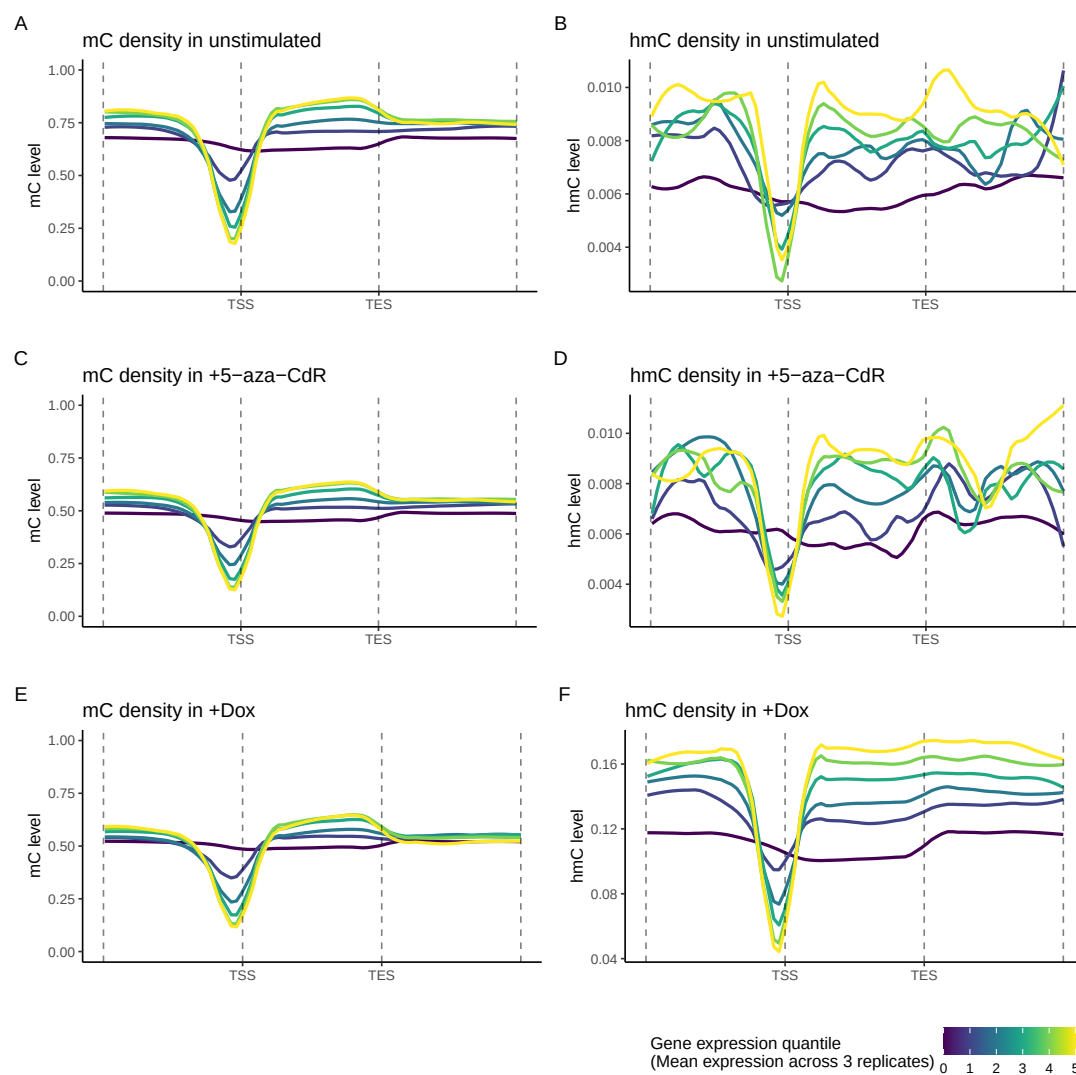


Figure 91: **mC and hmC gene density split by gene expression** Metagene plots summarising modification levels in genes in different expression quantiles. Those with zero counts are assigned quantile 0, the others are evenly split between the remaining 5 groups.

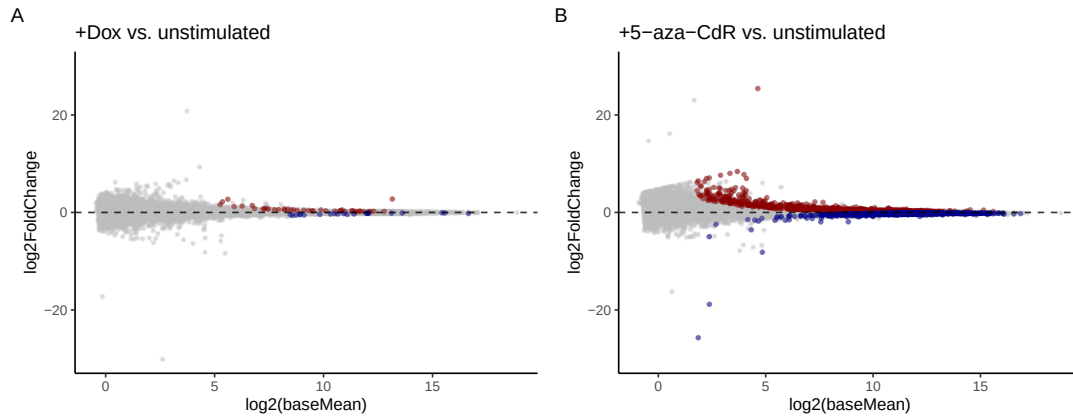


Figure 92: **MA plots.** MA plots comparing the mean expression measured by baseMean and the magnitudes of the log2FCs. Significant ($p\text{-adj} < 0.05$) changes are highlighted, with upregulated genes in red and downregulated in blue. A-B. Show the +Dox vs unstimulated and +5-aza-CdR vs unstimulated comparisons, respectively.

after the demethylating treatments. The relationship between hmC density at the promoter and expression is more complex, as higher expression change is associated with both hmC gain and uC gain at the promoter (Figure 94).

To check whether high magnitude changes after +Dox treatment are preferentially associated with uC gain or with hmC gain (since the treatment induces both), we looked for associations between the dominant promoter changes and the directionality and magnitude of expression changes (Figure ??). Based on the limited number of significant DEGs with sufficient uC gain and hmC gain signal in the promoter, we found that 6/7 of the most significantly altered genes had more hmC gain than uC gain, however clearly the number of genes in this group that passed the gain filters was extremely low. Reducing to a p-value cut-off, we saw that uC gain was more likely to cause upregulation compared to hmC gain, with 78.6% and 56.9% of genes upregulated in each group⁸⁹. This suggests that C gain at the promoter may be more transcription promoting than hmC gain, but with such low gene counts this conclusion is not well grounded.

Taking a slightly different approach to the same question, we selected genes similar

⁸⁹58.6% and 51.4% when considering all genes with sufficient hmC and uC gains.

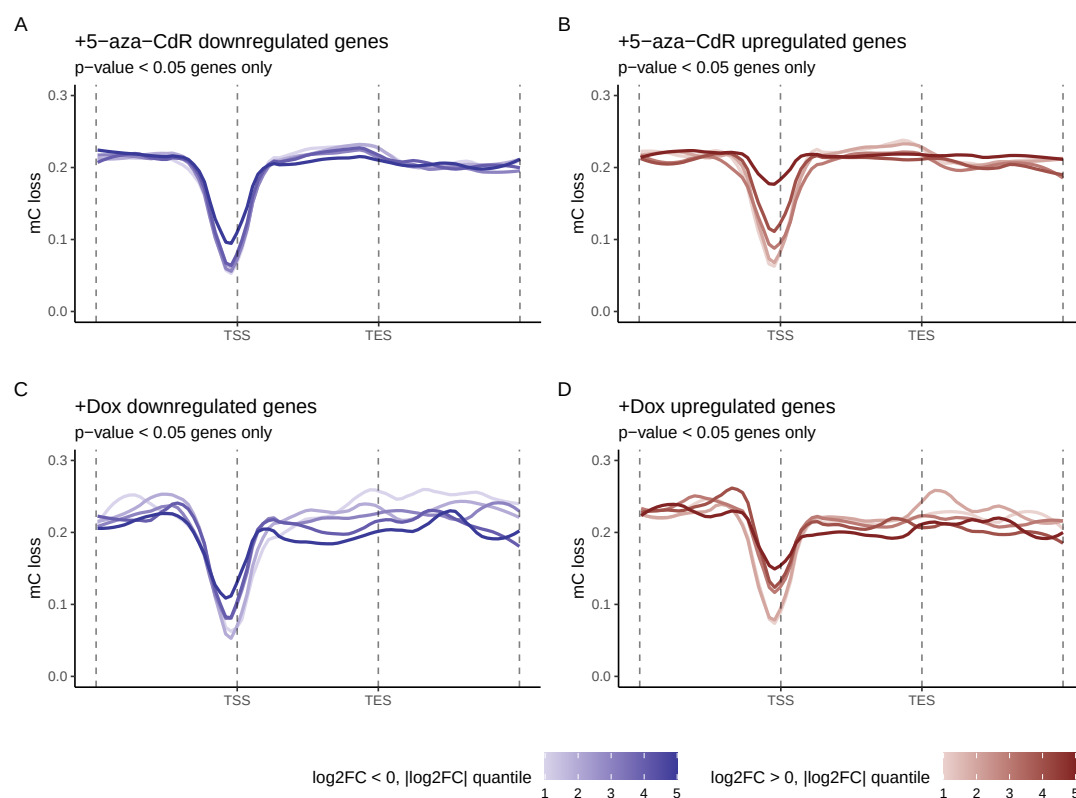


Figure 93: **mC loss gene density split by gene expression changes** Metagene plots summarising mC loss in genes with p-value < 0.05 in different log2FC quantiles, split by directionality of change.

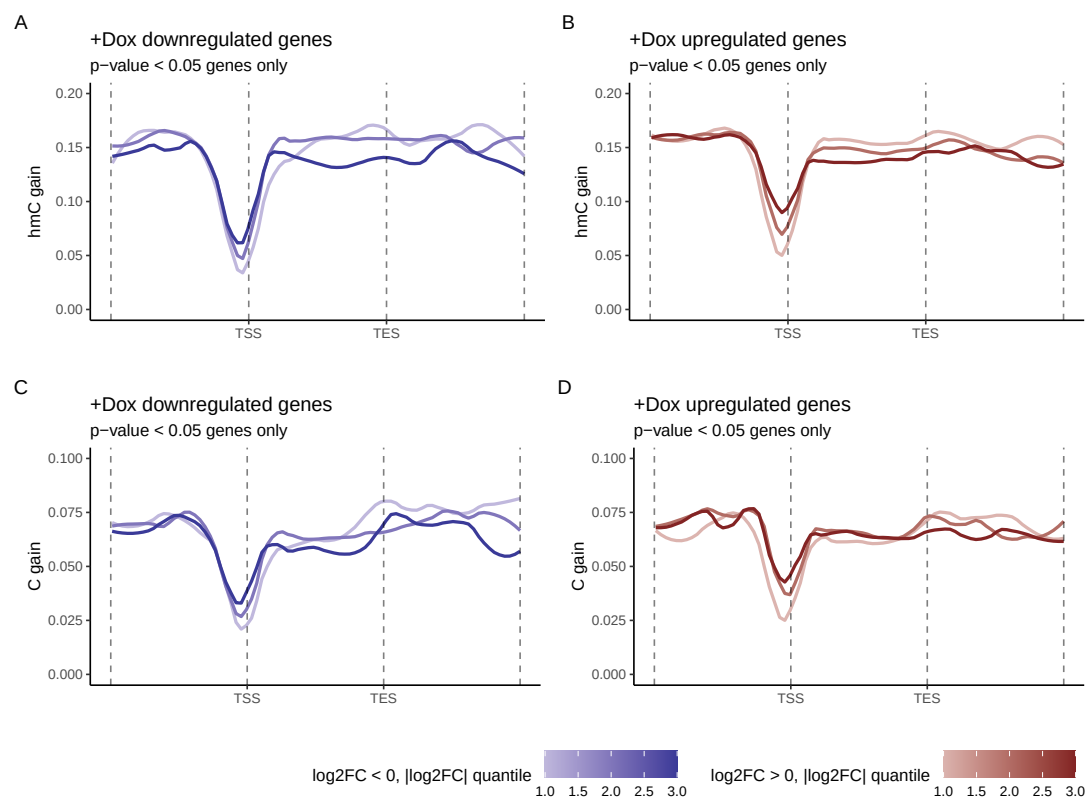


Figure 94: **hmC and unmodified C (uC) gain gene density split by gene expression changes** Metagene plots summarising hmC and uC gain in genes with p-value < 0.05 in different log2FC quantiles, split by directionality of change.

promoter-proximal uC gain in +5-aza-CdR as hmC gain in +Dox (with minimal uC gain in +Dox) (Figure 95C), and contrasted the effects on gene expression⁹⁰. We found 25 genes with stronger changes after +Dox (hmC gain) and 26 genes with stronger changes after +5-aza-CdR (uC gain), but most of the hmC gain genes didn't show strong changes in gene expression after either treatment (Figure 95D). Therefore uC does appear to be more permissive based on this small group of genes, consistent with the previous analysis⁹¹.

In summary, we found that the most upregulated (based on log2FC) genes after both treatments showed comparatively higher losses in TSS-proximal mC, but less change in the GBs. The TSS-proximal mC differences were less stark for downregulated genes, but the patterns were the same. This suggests that substantial TSS-proximal losses in mC are indicative of higher magnitude gains in gene expression. We found that uC gain appeared to be more permissive than hmC gain in terms of gene expression upregulation, however the gene counts in each group were very low and therefore this hypothesis will require further analysis (preferably in a cell line with minimal uC gain after +Dox treatment).

5.4 Results: Combining mRNA abundance and RNA polymerase occupancy data

In order to see whether the changes in nascent transcription that we saw in the previous chapter are reflected in mRNA, here we compare mRNA abundance and RNAP occupancy data directly.

5.4.1 Baseline PRO-Seq as a predictor of baseline mRNA

In this section we would like to see how informative various transcription metrics are in terms of predicting mRNA levels. In Section 4.2.10 we defined data-driven TSSs for each active gene, and subsequently specified elongation regions which were used to calculate other quantitative metrics such as pausing index (PI) and uniformity index

⁹⁰In simple terms, we select genes G with uC gain u after +5-aza-CdR treatment and hmC gain h after +Dox, such that $u \approx h$ and uC gain in +Dox < 0.05 .

⁹¹With the strong caveat that no promoter-methylation-independent effects are affecting these genes in each group.

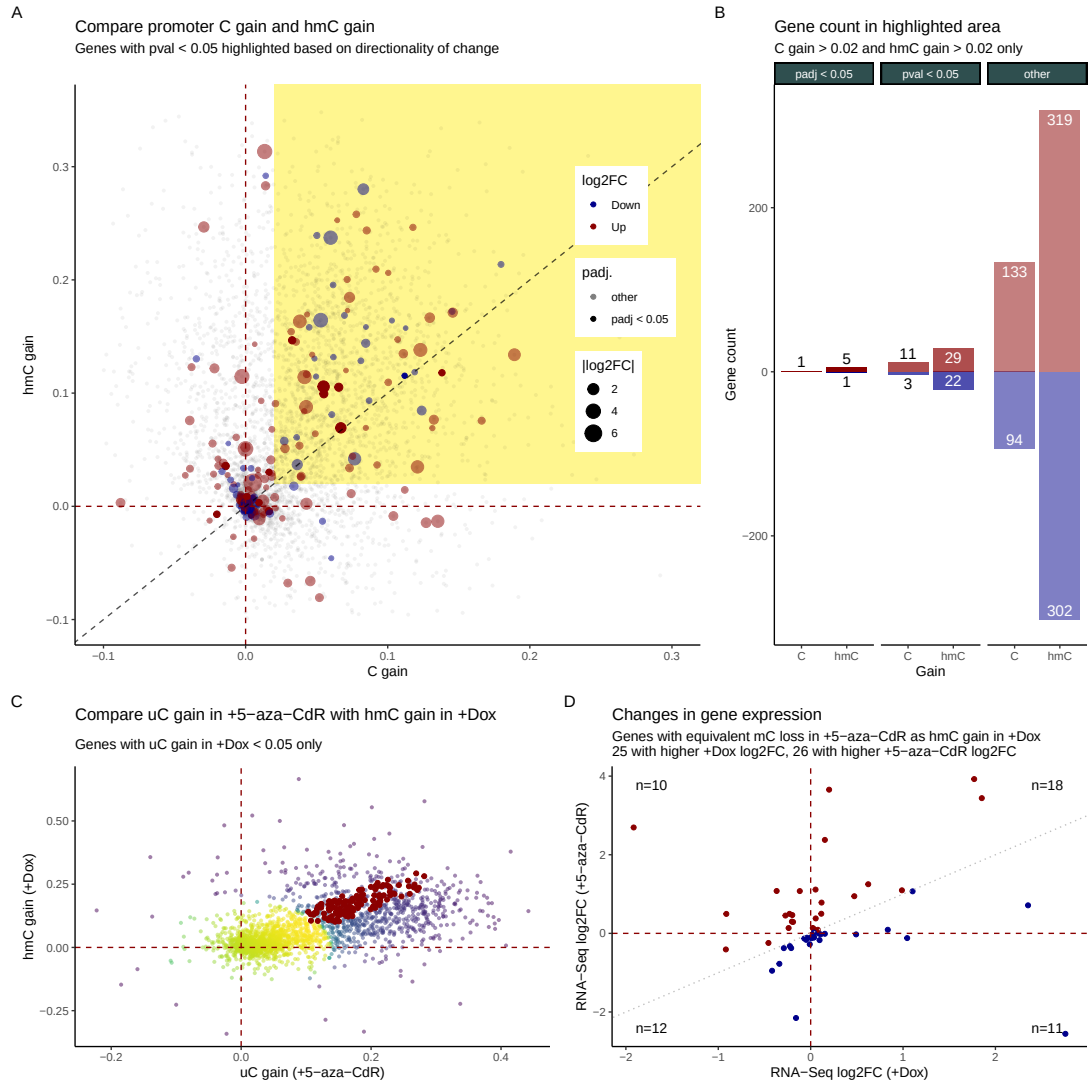


Figure 95: C gain vs hmC gain in promoters after +Dox treatment A. Scatterplot comparing average C and hmC gain in gene promoters (defined as [TSS-100,TSS+500]), with the genes showing differential expression after +Dox treatment highlighted based on directionality and size of change. Those with adjusted p-value < 0.05 are most opaque and genes with mean C gain and hmC gain exceeding 0.02 are highlighted in yellow. B. Comparing the number of genes with mostly C gain against those with dominant hmC gain, only considering genes falling in the yellow region. C. For genes with promoter-proximal uC gain < 0.05 after +Dox treatment, we compare the hmC gain after +Dox with uC gain after +5-aza-CdR and highlight the genes with similar uC and hmC gains. D. For the highlighted genes in C, we plot the log2FC in expression after each treatment. Points are coloured based on position relative to a line of slope 1 (where the log2FCs are equal after each treatment).

(UI). What is the relationship between the magnitudes of these quantities and mRNA expression, e.g. are high levels of pausing more common in highly transcribed genes?

We first looked at TSS-proximal sense initiation, which showed that highly expressed genes often had evidence of both high and low initiation peaks, but by partitioning those into major and minor initiation sites⁹² we see that the left density disappears and a slightly stronger linear relationship is evident (Figures 96A and B). The relationship is more complex for antisense peaks, highly expressed genes show a wide range of antisense initiation suggesting that it would be difficult to predict a gene's mRNA abundance based on the prevalence of antisense transcription, using the most abundant antisense peak results in better prediction but with slight horizontal smear in cases of similar mRNA abundance but many possible values of initiation signal (Figures 96C and D). Lastly, gene body transcription is generally low and the few genes with very high mRNA levels have minimal gene body initiation (Figures 96E and F). Therefore using the PRO-Cap initiation data for prediction is difficult.

The relationship with elongation was more linear, meaning that the abundance of PRO-Seq signal in the elongation regions was a better predictor of mRNA abundance i.e. nascent RNA processing does not “turn off” a substantial proportion of genes⁹³ (Figure 97A). Pausing index was not a strong predictor of mRNA, although the slight negative trend would suggest that highly expressed genes pause less at the TSS (Figure 97B) and indeed, the few genes with low uniformity, i.e. lots of gene-body pausing or initiation, were generally lowly expressed (Figure 97C).

5.4.2 Are changes in PRO-Seq predictive of changes in mRNA?

We noted some changes in RNAP occupancy after +5-aza-CdR and +Dox treatment in the previous chapter, and since we did confirm a relationship between expression and PRO-Seq abundance in the elongation regions (above), we hypothesised that some of these RNAP changes did manifest in altered mRNA.

We first looked at the changes after +5-aza-CdR treatment, and by comparing the log2FC in expression and elongation directly we found that high magnitude gains in

⁹²Simply using the TSS with highest count.

⁹³Which would result in high PRO-Seq but little mRNA, or a vertical smear.

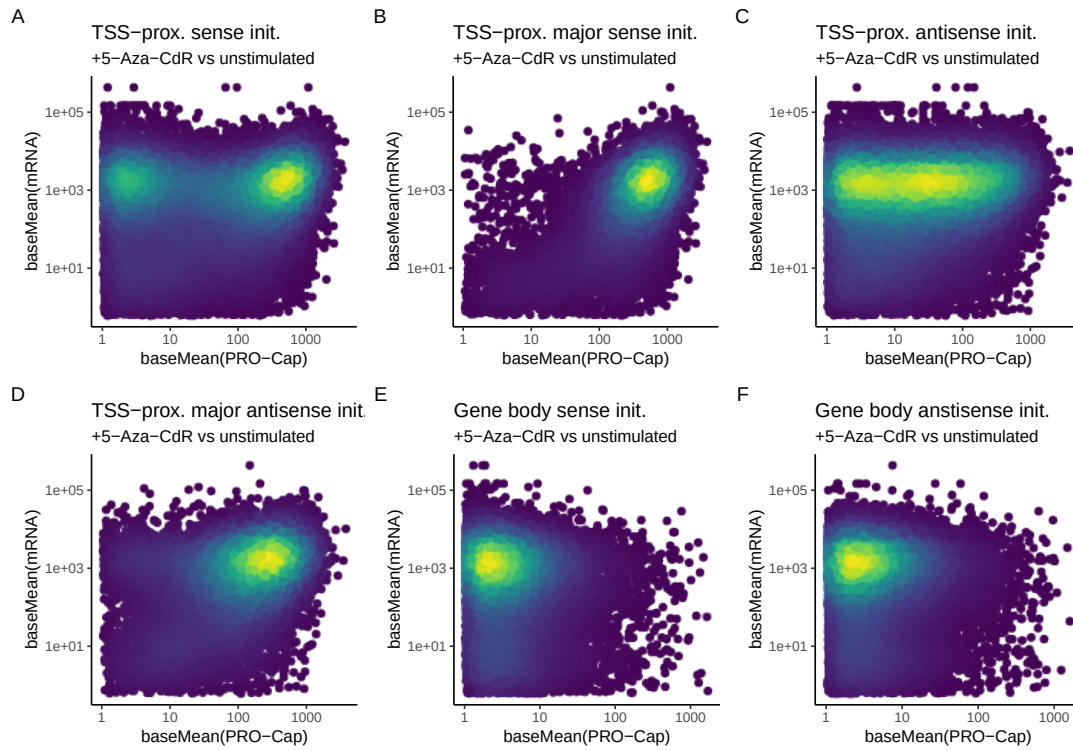


Figure 96: **Baseline mRNA versus initiation analysis.** Density scatter plots showing the baseMean of mRNA for each gene (from the +5-aza-CdR versus unstimulated analysis) compared to: A. TSS-proximal sense initiation (PRO-Cap). B. TSS-proximal sense initiation with only the highest count initiation peak considered for each gene (PRO-Cap). C. As A but antisense peaks only. D. As B but antisense peaks only. E-F. Gene body sense and antisense initiation (PRO-Cap).

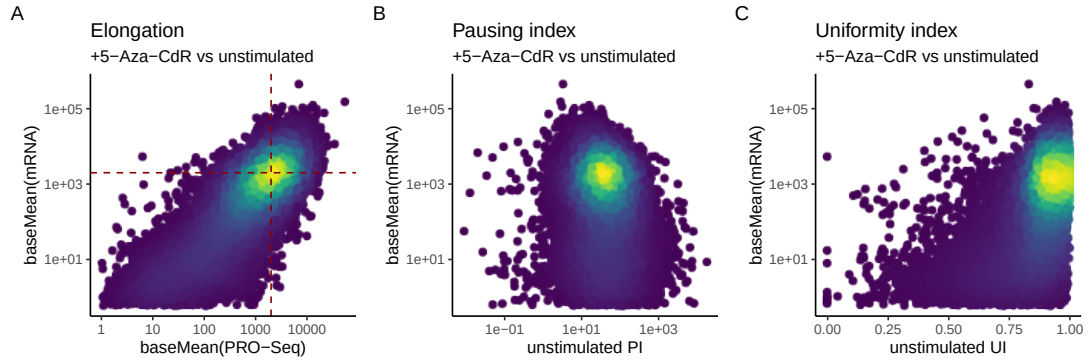


Figure 97: **Baseline mRNA analysis.** Density scatter plots showing the baseMean of mRNA for each gene (from the +5-aza-CdR versus unstimulated analysis) compared to: A. Elongation (PRO-Seq), with 2kb highlighted. B. Pausing index of the unstimulated sample (Section 4.2.11.2.2). C. Uniformity index of the unstimulated sample (see Section 4.2.11.2.3).

elongation usually resulted in gains in expression (Figure 98A, see examples in Figure 99). This relationship was even stronger when focussing on genes with significant ($p\text{-adj} < 0.05$) changes in both (Figure 98B). Although the +Dox treatment was less disruptive to both expression and elongation, the positive relationship does still hold for the 19 genes with changes (Figure 98C and D).

We then loosened the criteria to significant changes in mRNA *or* elongation, and saw that mRNA wasn't always a good predictor of elongation, particularly in the +5-aza-CdR case (Figures 100A and C). However positive changes in PRO-Seq correlated well with positive changes in mRNA (Figures 100B and D), suggesting that most of the PRO-Seq gains were the result of higher transcription rather than increased pausing. In contrast, negative changes in PRO-Seq were associated with both gain and loss in expression, however recall that PRO-Seq signal can be lower due to either lower expression or faster transcription (which can result in higher expression).

Assuming that higher mC abundance would reduce RNAP speed, we compared how positive and negative elongation changes related to mRNA with the addition of elongation region mC loss information. In +5-aza-CdR treated samples we noted that most expression increases are associated with high mC loss in the elongation region. Curiously, the genes with lower PRO-Seq abundance and mRNA after +5-aza-CdR

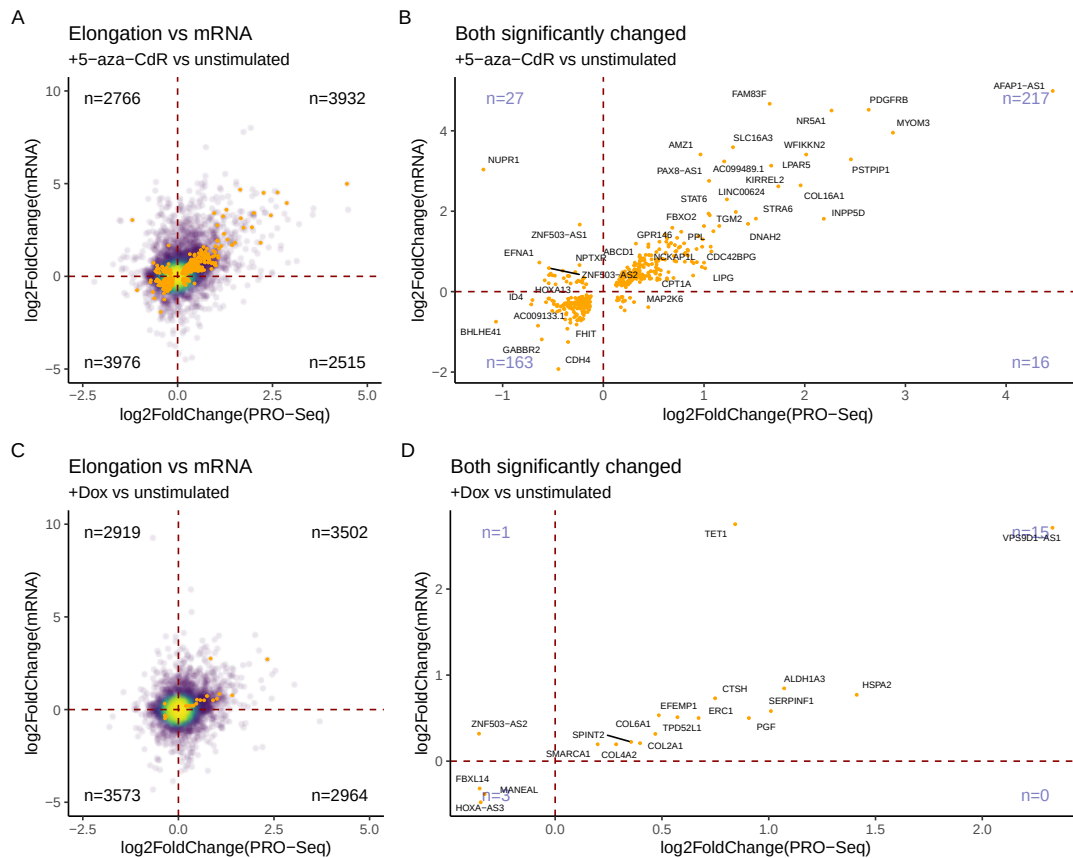


Figure 98: **Changes in mRNA and elongation.** A. Density scatter plot comparing the log2FC of elongation with log2FC of mRNA after +5-aza-CdR treatment. The genes with significant ($p\text{-adj} < 0.05$) are highlighted in orange. B. Scatter plot showing the significant genes from A. C-D. As A and B, but for +Dox versus unstimulated.

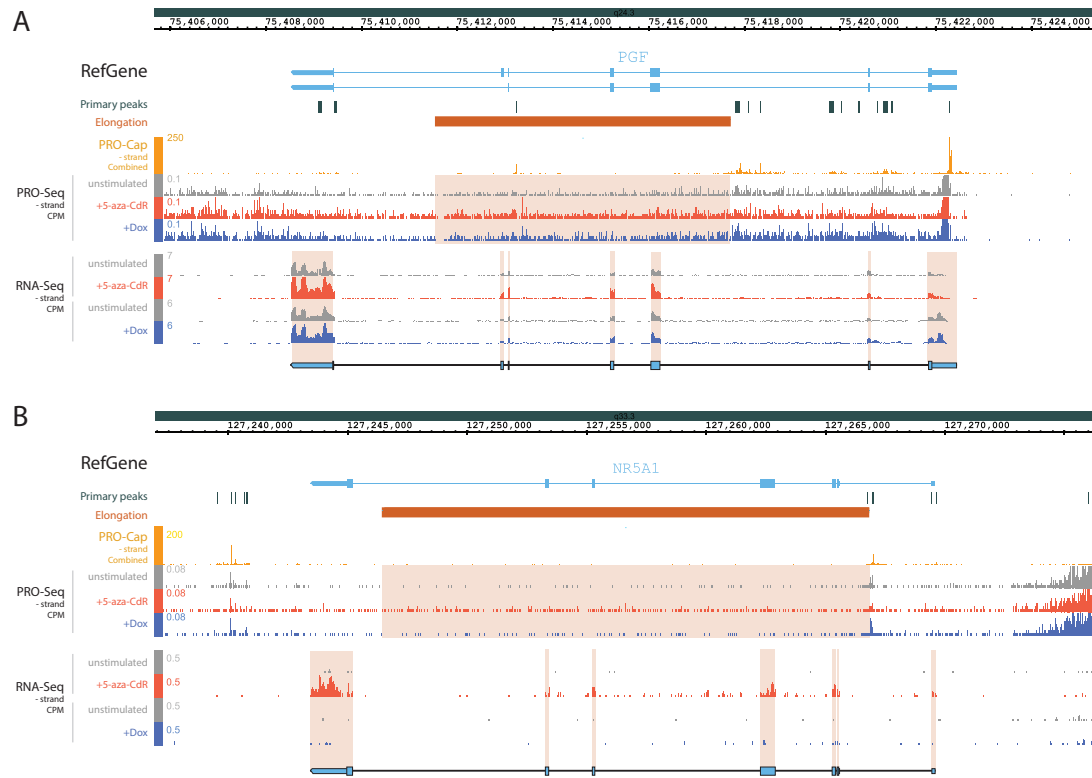


Figure 99: Examples of genes with differential elongation and mRNA signal. A-B. Genome browser views of PGF and NR5A1, respectively, with tracks showing combined strand-specific PRO-Cap signal (RNAP initiation), strand-specific PRO-Seq (RNAP occupancy) for each condition, and RNA-Seq abundances for each condition with each experiment group (CPM normalised in each case). The predicted TSSs (Primary peaks) and elongation regions are included. The region of interest within each experiment group is highlighted: elongation for PRO-Seq and exons for RNA-Seq.

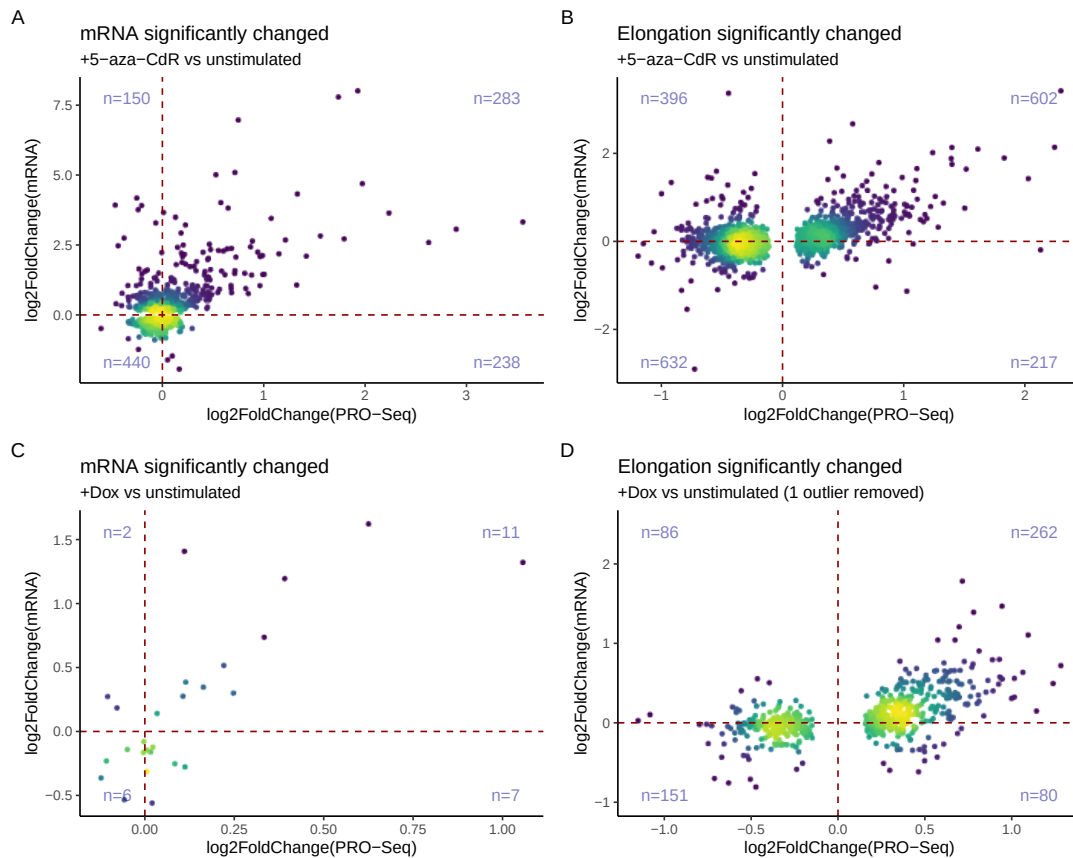


Figure 100: **Changes in mRNA or elongation.** A. Density scatter plot comparing the log2FC of elongation with log2FC of mRNA after +5-aza-CdR treatment. The genes with significant ($p\text{-adj} < 0.05$) are highlighted in orange. B. Scatter plot showing the significant genes from A. C-D. As A and B, but for +Dox versus unstimulated.

treatment seemed to lose **more** mC, i.e. in these cases high levels of demethylation resulted in both lower transcription and lower mRNA levels, suggesting that too much demethylation may have a negative effect on transcription (Figures 101A and B). Again, the +Dox treated group didn't show a strong relationship either way, for either mC loss or hmC gain (not shown) (Figures 101C and D).

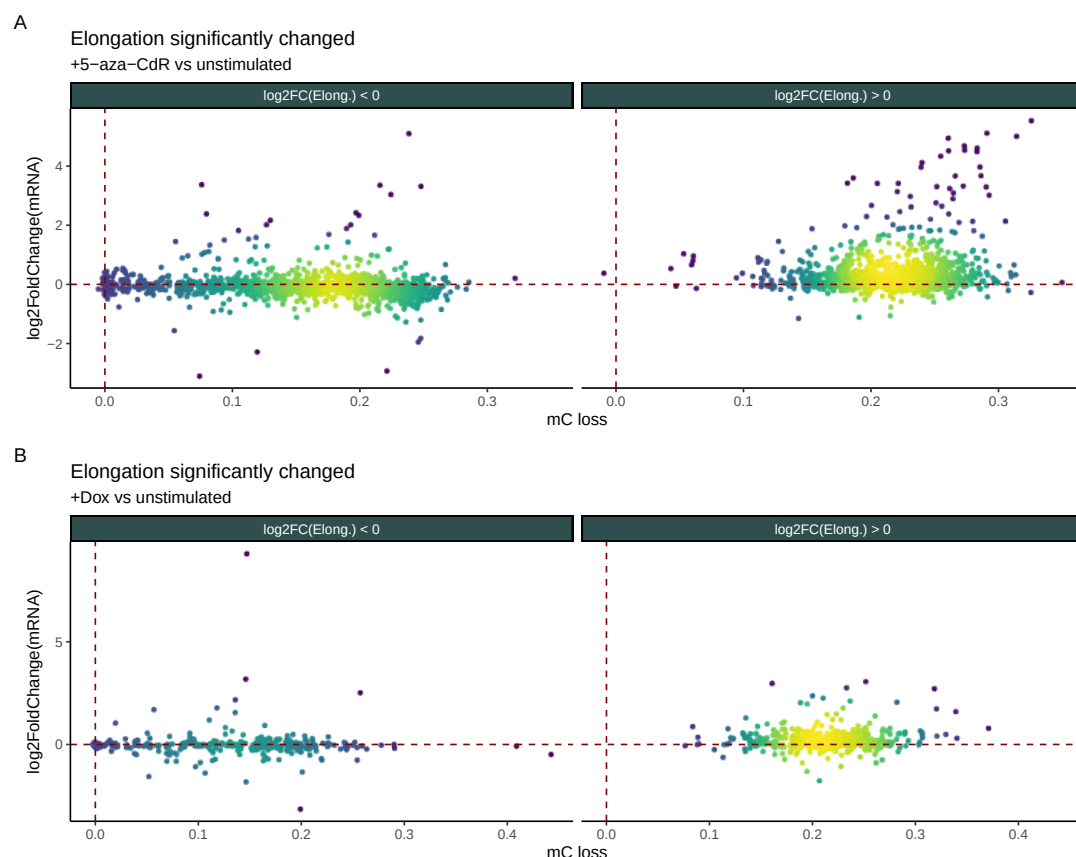


Figure 101: **Relationship between elongation, mRNA, and mC levels.** A-B. Density scatter plot comparing the mC loss after +5-aza-CdR treatment with log2FC of mRNA for genes with significant ($p\text{-adj} < 0.05$) elongation gains and losses, respectively. C-D. As A-B but for the +Dox versus unstimulated comparison.

To summarise, we noted that the significant changes we calculated in RNAP occupancy and mRNA abundance show concordance, i.e. genes with higher elongation signal tend to have higher mRNA. However the relationship isn't linear when non-significant changes are considered⁹⁴. In contrast, predicting mRNA change from elongation change

⁹⁴For example, predicting nascent RNA abundance from RNA-Seq data is difficult, with the

(upregulated only) was more feasible. This confirmed that, by looking at the occupancy data we gain more information about the effects of DNA modifications, since if occupancy was directly proportional to mRNA, then it would be sufficient to have mRNA data only. Lastly we confirmed that high mC loss was a common feature of genes with the most upregulation in both experiments.

5.5 Discussion

In this chapter we evaluated how mRNA levels are affected by DNA modification changes. We linked our observations to the previous chapters to understand the relationship between the bp-specific methylation changes induced, changes in occupancy of active RNAP, and gene expression changes.

In section 5.2.2 we evaluated the possibility of methylation-independent changes induced by each treatment. We found few changes in expression upon inducing catalytically dead TET1-CD, suggesting that the changes we did note after TET1-CD induction are indeed related to its enzymatic activity. The effects of DNA-protein crosslinks induced by adding 5-aza-CdR were mimicked by adding 4-thio-2'-deoxythymidine and UVA light such that the resulting protein-DNA complex abundance was similar to that post 5-aza-CdR treatment. Although the overlap was significant (both for upregulated and downregulated genes), most of the +5-aza-CdR induced changes (87% of up- and 89% of downregulated) were cross-linking independent.

Treatment with 5-aza-CdR was significantly more deregulating than TET1-CD upregulation, with 1,982 and 64 DEGs in HEK293 cells, respectively⁹⁵. Demethylation of HEK293 cells with 5-aza-CdR resulted in 903 up- and 1079 downregulated genes, where the highest magnitude changes were upregulating. In their study, Yang et al. [266] noted that 5-aza-CdR treatment both reactivated some genes but also decreased overexpression of others, which is in agreement with these findings. In contrast, TET1-CD upregulation appears to have little effect on gene expression, with only 42 up- and 22 down-regulated genes in HEK293 cells. This is consistent with a previous study of exception of high magnitude upregulation usually being associated with high magnitude changes in elongation.

⁹⁵2,487 (+5-aza-CdR) and 902 (+Dox) DEGs in U2OS.

HEK293T cells overexpressing a CXXC-less isoform of TET1 which only showed 23 up- and 22 down-regulated genes [78]. Despite the differences in numbers, there was still substantial overlap between treatments, with 57.1% of genes significantly upregulated by +Dox also upregulated by +5-aza-CdR (overlap for downregulated was 27.3%)⁹⁶.

Treatment of U2OS cells with +Dox resulted in 513 up- and 389 downregulated genes. This showed that hmC induction does have the potential to cause greater changes in other cell lines, with 316 genes that were affected by +Dox but not by +5-aza-CdR. Unfortunately without DNA modification and RNAP occupancy data for this cell line we could not check whether the changes were the consequence of altered transcription. However, this confirmed that using an alternative cell line could result in more DEGs upon treatment, and, as mentioned in previous chapters, using multiple cell lines with varying levels of mC (particularly at the promoter) would provide a more complete picture of the effects of DNA modifications.

Similarly to other groups, we confirmed that both TSS-proximal and gene body DNA modifications are related to expression levels in HEK293 cells, with the most highly expressed genes having minimal TSS-proximal mC and hmC and highest rates of gene body modifications. The most altered genes lost more TSS-proximal mC on average, particularly those that were upregulated. This is not surprising as promoter-proximal mC is associated with repression, thus promoter-mC-repressed genes were more likely to be affected since promoters could be directly altered by the demethylating treatments.

In contrast, gene body methylation was less changed in these cases, stemming from initial low expression with correspondingly low expected baseline methylation. When we compared hmC gain and unmodified C gain at promoters, we found trends suggesting that uC gain was more associated with gene upregulation than downregulation. However, there were few genes for which we could contrast the effects of promoter-proximal hmC gain and uC gain on gene expression directly. To increase the pool of comparable genes, a cell-cycle arrested system with reduced uC gain after TET1-CD upregulation could be used (see Section 3.3, and Figure 39).

We saw a correlation (Spearman's $\rho = 0.74$) between elongation signal (PRO-Seq)

⁹⁶67.8% (upregulated) and 61.6% (downregulated) of DEGs in U2OS.

and expression (RNA-Seq). This was comparable to another study comparing gene expression and PRO-Seq elongation [21] and suggested that baseline elongating RNAP abundance is indeed predictive of gene expression. Some of the variation can be explained by the extra processing steps required to convert nascent RNA to mRNA as well as degradation. The comparison of treatment induced changes in elongation and mRNA confirmed a linear relationship (it was rare to see a significant increase in elongation but decrease in mRNA), however without elongation data we would not be able to accurately predict elongation changes from mRNA changes. This confirmed that, despite the strong relationship between baseline mRNA and nascent transcription, the effects of the perturbations were difficult to evaluate without both sets of data.

In summary, we concluded that the hallmark of the most upregulated genes, for which both transcription and mRNA was usually affected, was significant mC loss in the elongation region. Together with the pausing analysis, we found that mC loss results in gains in RNAP pause-release and expression, so why would highly expressed genes be hypermethylated in the gene bodies? One hypothesis relates to the trade-off between speed and fidelity, if faster RNAPs make more mistakes then genes which are most expressed (and contribute most to the pool of mRNA) should have more mC “speedbumps”. Thus far, the relationship between RNAP elongation rate and fidelity has not been thoroughly described. However, in a study of transcription dynamics, Depken et al. [48] proposed a model whereby intermittent pausing during transcription, which allows for required proofreading and backtracking to take place, is used as a method to mediate between speed and fidelity. Similarly, in the context of viral genetic drift, Fitzsimmons et al. [68] described a model whereby high mutation rates are a consequence of selection for faster replicative speed. Another factor to consider are the mutagenic properties of DNA modifications. In their study of mutations at mC and hmC in brain, Tomkova et al. [239] found that hmC is less mutagenic than C and mC, meaning that accumulation of hmC in gene bodies may also play an independent protective role.

6 Discussion of future directions

We have used the chapter discussions to expand on some changes that could be made to the system retrospectively. Here we consider all our results and discuss further changes to experimental design, additional experiments, as well as related questions.

6.1 Changes to experimental design

Throughout this project we have seen few high-magnitude changes evident in HEK293 cells after demethylation with +5-aza-CdR. In contrast, treating U2OS cells resulted in almost a 10-fold increase in the number of genes with significant changes after TET1-CD upregulation only, confirming that the level of deregulation can vary significantly between cell lines. To ensure we give the system more opportunity to show changes, I believe that an alternative cell line should be chosen based on its baseline methylation which would allow more genes to lose methylation. A highly methylated cell line can be created through impairment of active demethylation. In IDH mutant cell lines, TET cannot oxidise mC to hmC due to an accumulation of 2-hydroxyglutarate (2-HG) and reduced $\alpha - KG$, as a result there is an accumulation of mC in the genome [67, 206, 12, 20]. Apart from an mC upregulation, the lack of hmC in these cell lines can make the effect of hmC induction on transcription more clear.

We found that the changes induced by TET1-CD upregulation were weaker than those induced by +5-aza-CdR, but due to the unmodified C gain after TET1-CD the hmC-specific effect wasn't very clear. We did note a higher level of permissiveness, namely it's better to have uC than hmC at the promoter for expression upregulation, but the finding was based on very few genes. In order to study the hmC-specific changes, a system where all mC loss is accounted for by hmC gain is preferable. Dr. Hiromi Tagoh used a system of cell cycle arrest to reduce the effects of replication-associated passive demethylation and found very promising results for U2OS cells (Section 3.3). Using this system, it would be interesting to calculate the proportion of overlaps between differentially expressed genes and contrast that to our current results. We could see whether there is significant overlap, and elucidate whether the overlaps we noted in our analyses were the consequence of uC gain in both or if hmC gain and uC gain had similar

effects on some genes. With the above taken into account, an experimental design with increasing hmC gains could also be beneficial in refining the role of hmC. By using the same concentrations of Dox, but extracting the cells at different time points (e.g. 1 hour, 2 hours, 4 hours, 8 hours, 12 hours), we could see how the progressive change from mC to hmC affects gene expression over time (Figure 102).

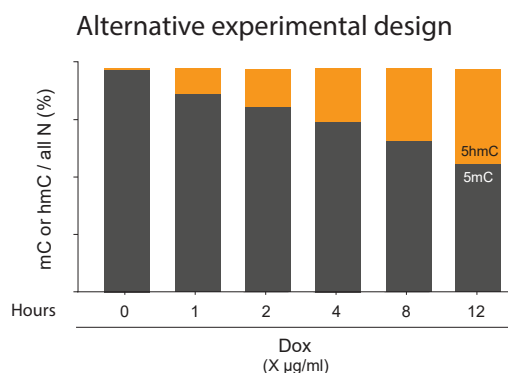


Figure 102: **Simulated ideal alternative hmC induction.** By harvesting the cells at different time-points, the effects of hmC can be seen more clearly.

We incorporated some controls to check for DNA modification-independent effects of 5-aza-CdR on transcription, and found that 1,241 genes were differentially expressed when DNA-protein crosslinks were induced. Although the overlap with the +5-aza-CdR DEGs wasn't big, 13.2% of up- and 11.1% of downregulated, it was still significant. Thus, an alternative method to reduce mC genome-wide would be preferred. Alternative possible inducible systems to reduce DNMT1 could include the use of shRNA [29], knockout-rescue CRISPR [248], a REMOTE⁹⁷-control system [129], or Cre recombinase [59]. All of these (with maybe the exception of shRNA) are more complex than using 5-aza-CdR (which has been used as a demethylase in numerous studies), and would need to be tested for off-target effects.

⁹⁷Reversible Manipulation of Transcription at Endogenous loci.

6.2 Alternative and additional experiments

PRO-Seq does not differentiate between a slow RNAP and many fast RNAPs since the densities would be similar. Adding a temporal component to the analysis would improve interpretation and detection of changes in RNAP speed. One option is time variant PRO-Seq (TV-PRO-Seq) [270], where multiple PRO-Seq experiments are prepared, and since each has a different run-on time they can be used to estimate pause-release through tracking the order in which positions incorporate a biotinylated base (paused RNAPs will be mapped more in longer duration run-on groups).

Although we saw changes in intergenic initiation peaks which may be distal enhancers, without an accurate map of enhancers it was difficult to link any gene expression changes with these distal elements. To evaluate how DNA modifications affect enhancer regions, an accurate map of enhancer annotations and enhancer-gene interaction is required. Although nascent RNA production can be used to accurately map enhancers [93], in order to annotate enhancers and their state (active, primed, poised), we would need to map histone modifications in each of the samples (control, mC loss, hmC gain). This would allow us to detect which enhancers may become active, and, using additional Hi-C data, identify the genes whose expression is dependent on these enhancers.

6.3 Related questions

With the above improvements to experimental design and methods used, we can re-address some questions that were not included in this thesis for brevity. We did not find any strong effects on splicing in our data, but links have been proposed between pausing, DNA modifications, and splicing. Multiple studies have linked DNA modifications to alternative splicing [118, 163, 256, 218, 146, 15]. Similarly, by using RNAPII rate mutants, Fong et al. [69] found that slow and fast elongation rates could result in intron retention and exon inclusion, hypothesising that an optimal rate of transcriptional rate is required for normal splicing. Thus, a deeper analysis of splicing and intron retention using a more optimised set of samples would be interesting to look at.

We noted the possible link between speed and fidelity in the previous chapter. Un-

fortunately, significant mutagenesis within the 48 hour treatment window is simply not long enough to detect changes. To understand the relationship, we could compare the mutation rate within genes, split by their baseline elongation rate (since elongation rates usually vary between genes [46, 244, 81]). To deconvolute speed from methylation, however, an organism or cell line with minimal methylation should be used. The extremely low levels of mC in *Drosophila* aren't detectable by bisulfite sequencing [49], thus this could be a candidate for comparing the relationship between RNAP speed and mutagenesis. Preliminary analyses of elongation rates should be carried out first to ensure that we have well-defined elongation rate gene groups, particularly since *Drosophila* embryos showed less than 25% variation between genic elongation rates Fukaya et al. [72].

7 Conclusion

In this thesis we have shown that DNA modifications affect multiple states of active transcription, as measured by occupancy of actively transcribing RNAP.

We found modest changes in initiation, pausing, and elongation after both demethylation with +5-aza-CdR and hmC induction through TET1-CD upregulation. The highest magnitude changes suggested that DNA modifications play an inhibitory and pausing role since both mC-reducing treatments were associated with less pausing (both at the TSS and in the gene body) and higher abundance of elongating RNAP.

The high-magnitude positive changes in active transcription were reflected in mRNA abundance, suggesting that DNA modifications can regulate transcription through their interaction with actively elongating RNAP.

Demethylation was consistently more disruptive compared to hmC gain. This indicates that the role of hmC isn't transcription-promoting compared to unmodified C, and therefore the abundance of hmC in gene bodies is more likely to be related to other factors, such as increased fidelity through slowing down RNAP.

References

- [1] E. A. Alcid and T. Tsukiyama. ATP-dependent chromatin remodeling shapes the long noncoding RNA landscape. *Genes & Development*, 28(21):2348–2360, nov 2014. ISSN 0890-9369. doi: 10.1101/gad.250902.114. URL <http://genesdev.cshlp.org/lookup/doi/10.1101/gad.250902.114>.
- [2] J. An, E. González-Avalos, A. Chawla, M. Jeong, I. F. López-Moyado, W. Li, M. A. Goodell, L. Chavez, M. Ko, and A. Rao. Acute loss of TET function results in aggressive myeloid cancer in mice. *Nature Communications*, 6(1):10071, dec 2015. ISSN 2041-1723. doi: 10.1038/ncomms10071. URL <http://www.nature.com/articles/ncomms10071>.
- [3] R. Andersson, C. Gebhard, I. Miguel-Escalada, I. Hoof, J. Bornholdt, M. Boyd, Y. Chen, X. Zhao, C. Schmidl, T. Suzuki, E. Ntini, E. Arner, E. Valen, K. Li, L. Schwarzfischer, D. Glatz, J. Raithel, B. Lilje, N. Rapin, F. O. Bagger, M. Jørgensen, P. R. Andersen, N. Bertin, O. Rackham, A. M. Burroughs, J. K. Baillie, Y. Ishizu, Y. Shimizu, E. Furuhashi, S. Maeda, Y. Negishi, C. J. Mungall, T. F. Meehan, T. Lassmann, M. Itoh, H. Kawaji, N. Kondo, J. Kawai, A. Lennartsson, C. O. Daub, P. Heutink, D. A. Hume, T. H. Jensen, H. Suzuki, Y. Hayashizaki, F. Müller, A. R. R. Forrest, P. Carninci, M. Rehli, and A. Sandelin. An atlas of active enhancers across human cell types and tissues. *Nature*, 507(7493):455–461, mar 2014. ISSN 0028-0836. doi: 10.1038/nature12787. URL <http://www.nature.com/articles/nature12787>.
- [4] R. Andersson, P. Refsing Andersen, E. Valen, L. J. Core, J. Bornholdt, M. Boyd, T. Heick Jensen, and A. Sandelin. Nuclear stability and transcriptional directionality separate functionally distinct RNA species. *Nature Communications*, 5(1):5336, dec 2014. ISSN 2041-1723. doi: 10.1038/ncomms6336. URL <http://www.nature.com/articles/ncomms6336>.
- [5] S. Andrews. FastQC: A quality control tool for high throughput sequence data, 2010. URL www.bioinformatics.babraham.ac.uk/projects/fastqc/.

- [6] M. Arechederra, F. Daian, A. Yim, S. K. Bazai, S. Richelme, R. Dono, A. J. Saurin, B. H. Habermann, and F. Maina. Hypermethylation of gene body CpG islands predicts high dosage of functional oncogenes in liver cancer. *Nature Communications*, 9(1):3164, dec 2018. ISSN 2041-1723. doi: 10.1038/s41467-018-05550-5. URL <http://www.nature.com/articles/s41467-018-05550-5>.
- [7] P. Arnaud, C. Goubely, T. Peelissier, and J.-M. Deragon. SINE Retroposons Can Be Used In Vivo as Nucleation Centers for De Novo Methylation. *Molecular and Cellular Biology*, 20(10):3434–3441, may 2000. ISSN 0270-7306. doi: 10.1128/MCB.20.10.3434-3441.2000. URL <https://journals.asm.org/doi/10.1128/MCB.20.10.3434-3441.2000>.
- [8] E. Arner, C. O. Daub, K. Vitting-Seerup, R. Andersson, B. Lilje, F. Drabløs, A. Lennartsson, M. Rönnerblad, O. Hrydziusko, M. Vitezic, T. C. Freeman, A. M. N. Alhendi, P. Arner, R. Axton, J. K. Baillie, A. Beckhouse, B. Bodega, J. Briggs, F. Brombacher, M. Davis, M. Detmar, A. Ehrlund, M. Endoh, A. Es-lami, M. Fagiolini, L. Fairbairn, G. J. Faulkner, C. Ferrai, M. E. Fisher, L. Forrester, D. Goldowitz, R. Guler, T. Ha, M. Hara, M. Herlyn, T. Ikawa, C. Kai, H. Kawamoto, L. M. Khachigian, S. P. Klinken, S. Kojima, H. Koseki, S. Klein, N. Mejhert, K. Miyaguchi, Y. Mizuno, M. Morimoto, K. J. Morris, C. Mummery, Y. Nakachi, S. Ogishima, M. Okada-Hatakeyama, Y. Okazaki, V. Orlando, D. Ovchinnikov, R. Passier, M. Patrikakis, A. Pombo, X.-Y. Qin, S. Roy, H. Sato, S. Savvi, A. Saxena, A. Schwegmann, D. Sugiyama, R. Swoboda, H. Tanaka, A. Tomoiu, L. N. Winteringham, E. Wolvetang, C. Yanagi-Mizuochi, M. Yoneda, S. Zabierowski, P. Zhang, I. Abugessaisa, N. Bertin, A. D. Diehl, S. Fukuda, M. Furuno, J. Harshbarger, A. Hasegawa, F. Hori, S. Ishikawa-Kato, Y. Ishizu, M. Itoh, T. Kawashima, M. Kojima, N. Kondo, M. Lizio, T. F. Meehan, C. J. Mungall, M. Murata, H. Nishiyori-Sueki, S. Sahin, S. Nagao-Sato, J. Severin, M. J. L. de Hoon, J. Kawai, T. Kasukawa, T. Lassmann, H. Suzuki, H. Kawaji, K. M. Summers, C. Wells, D. A. Hume, A. R. R. Forrest, A. Sandelin, P. Carninci, and Y. Hayashizaki. Transcribed enhancers lead

- waves of coordinated transcription in transitioning mammalian cells. *Science*, 347 (6225):1010–1014, feb 2015. ISSN 0036-8075. doi: 10.1126/science.1259418. URL <https://www.sciencemag.org/lookup/doi/10.1126/science.1259418>.
- [9] M. Bachman, S. Uribe-Lewis, X. Yang, M. Williams, A. Murrell, and S. Balasubramanian. 5-Hydroxymethylcytosine is a predominantly stable DNA modification. *Nature Chemistry*, 6(12):1049–1055, dec 2014. ISSN 1755-4330. doi: 10.1038/nchem.2064. URL <http://www.nature.com/articles/nchem.2064>.
- [10] C. B. Ball, K. A. Nilson, and D. H. Price. Use of the nuclear walk-on methodology to determine sites of RNA polymerase II initiation and pausing and quantify nascent RNAs in cells. *Methods*, 159-160(December 2018): 165–176, apr 2019. ISSN 10462023. doi: 10.1016/j.ymeth.2019.02.003. URL <https://doi.org/10.1016/j.ymeth.2019.02.003https://linkinghub.elsevier.com/retrieve/pii/S1046202318303098>.
- [11] E. Barbieri, C. Hill, M. Quesnel-Vallières, A. J. Zucco, Y. Barash, and A. Gardini. Rapid and Scalable Profiling of Nascent RNA with fastGRO. *Cell Reports*, 33 (6):108373, nov 2020. ISSN 22111247. doi: 10.1016/j.celrep.2020.108373. URL <https://linkinghub.elsevier.com/retrieve/pii/S2211124720313620>.
- [12] C. Bardella, O. Al-Dalahmah, D. Krell, P. Brazauskas, K. Al-Qahtani, M. Tomkova, J. Adam, S. Serres, H. Lockstone, L. Freeman-Mills, I. Pfeffer, N. Sibson, R. Goldin, B. Schuster-Böeckler, P. J. Pollard, T. Soga, J. S. McCullagh, C. J. Schofield, P. Mulholland, O. Ansorge, S. Kriaucionis, P. J. Ratcliffe, F. G. Szele, and I. Tomlinson. Expression of Idh1R132H in the Murine Subventricular Zone Stem Cell Niche Recapitulates Features of Early Gliomagenesis. *Cancer Cell*, 30(4):578–594, oct 2016. ISSN 15356108. doi: 10.1016/j.ccell.2016.08.017. URL <https://linkinghub.elsevier.com/retrieve/pii/S1535610816304020>.
- [13] K. R. Barnett, B. E. Decato, T. J. Scott, T. J. Hansen, B. Chen, J. Attalla, A. D. Smith, and E. Hodges. ATAC-Me Captures Prolonged DNA Methylation of Dynamic Chromatin Accessibility Loci during Cell Fate Transitions. *Molecular Cell*, 77(6):1350–1364.e6, mar 2020. ISSN 10972765. doi: 10.1016/j.molcel.

2020.01.004. URL <https://doi.org/10.1016/j.molcel.2020.01.004https://linkinghub.elsevier.com/retrieve/pii/S1097276520300046>.

- [14] C. R. Bartman, N. Hamagami, C. A. Keller, B. Giardine, R. C. Hardison, G. A. Blobel, and A. Raj. Transcriptional Burst Initiation and Polymerase Pause Release Are Key Control Points of Transcriptional Regulation. *Molecular Cell*, 2019. ISSN 10974164. doi: 10.1016/j.molcel.2018.11.004.
- [15] E. Batsche, J. Yi, O. Mauger, E. Kornobis, B. Hopkins, C. Hanmer-Lloyd, and C. Muchardt. CD44 alternative splicing senses intragenic DNA methylation in tumors via direct and indirect mechanisms. *Nucleic Acids Research*, 49(11):6213–6237, jun 2021. ISSN 0305-1048. doi: 10.1093/nar/gkab437. URL <https://academic.oup.com/nar/article/49/11/6213/6292109>.
- [16] T. Baubec, D. F. Colombo, C. Wirbelauer, J. Schmidt, L. Burger, A. R. Krebs, A. Akalin, and D. Schübeler. Genomic profiling of DNA methyltransferases reveals a role for DNMT3B in genic methylation. *Nature*, 520(7546):243–247, apr 2015. ISSN 0028-0836. doi: 10.1038/nature14176. URL <http://www.nature.com/articles/nature14176>.
- [17] M. Bi, Z. Zhang, P. Xue, K. Hernandez, H. Wang, X. Fu, C. D. Angelis, Z. Gao, J. Ruan, V. X. Jin, Q. Wang, E. Marangoni, T. H.-M. Huang, L. Chen, C. K. Glass, W. Li, R. Schiff, and Z. Liu. Coordinate Enhancer Reprogramming by GATA3 and AP1 Promotes Phenotypic Plasticity to Achieve Breast Cancer Endocrine Resistance. *bioRxiv*, 2019. doi: 10.1101/767871.
- [18] M. Bi, Z. Zhang, Y.-Z. Jiang, P. Xue, H. Wang, Z. Lai, X. Fu, C. De Angelis, Y. Gong, Z. Gao, J. Ruan, V. X. Jin, E. Marangoni, E. Montaudon, C. K. Glass, W. Li, T. H.-M. Huang, Z.-M. Shao, R. Schiff, L. Chen, and Z. Liu. Enhancer reprogramming driven by high-order assemblies of transcription factors promotes phenotypic plasticity and breast cancer endocrine resistance. *Nature Cell Biology*, 22(6):701–715, jun 2020. ISSN 1465-7392. doi: 10.1038/s41556-020-0514-z. URL <http://www.nature.com/articles/s41556-020-0514-z>.

- [19] K. Bian, S. A. P. Lenz, Q. Tang, F. Chen, R. Qi, M. Jost, C. L. Drennan, J. M. Essigmann, S. D. Wetmore, and D. Li. DNA repair enzymes ALKBH2, ALKBH3, and AlkB oxidize 5-methylcytosine to 5-hydroxymethylcytosine, 5-formylcytosine and 5-carboxylcytosine in vitro. *Nucleic Acids Research*, 47(11):5522–5529, jun 2019. ISSN 0305-1048. doi: 10.1093/nar/gkz395. URL <https://academic.oup.com/nar/advance-article/doi/10.1093/nar/gkz395/5494752https://academic.oup.com/nar/article/47/11/5522/5494752>.
- [20] R. Blede, V. Vasudevaraja, S. Patel, J. Stafford, J. Serrano, G. Esposito, L. M. Tredwin, N. Goodman, A. Kloetgen, J. G. Golfinos, D. Zagzag, B. Weigelt, A. J. Iafrate, E. P. Sulman, A. S. Chi, S. Dogan, J. S. Reis-Filho, S. Chiang, D. Placantonakis, A. Tsigos, and M. Snuderl. Functional and topographic effects on DNA methylation in IDH1/2 mutant cancers. *Scientific Reports*, 9(1):16830, dec 2019. ISSN 2045-2322. doi: 10.1038/s41598-019-53262-7. URL <http://www.nature.com/articles/s41598-019-53262-7>.
- [21] A. Blumberg, Y. Zhao, Y.-F. Huang, N. Dukler, E. J. Rice, A. G. Chivu, K. Krumholz, C. G. Danko, and A. Siepel. Characterizing RNA stability genome-wide through combined analysis of PRO-seq and RNA-seq data. *BMC Biology*, 19(1):30, dec 2021. ISSN 1741-7007. doi: 10.1186/s12915-021-00949-x. URL <https://bmcbiol.biomedcentral.com/articles/10.1186/s12915-021-00949-x>.
- [22] C. Bock, I. Beerman, W.-H. Lien, Z. D. Smith, H. Gu, P. Boyle, A. Gnirke, E. Fuchs, D. J. Rossi, and A. Meissner. DNA Methylation Dynamics during In Vivo Differentiation of Blood and Skin Stem Cells. *Molecular Cell*, 47(4):633–647, aug 2012. ISSN 10972765. doi: 10.1016/j.molcel.2012.06.019. URL <https://linkinghub.elsevier.com/retrieve/pii/S1097276512005448>.
- [23] G. T. Booth, I. X. Wang, V. G. Cheung, and J. T. Lis. Divergence of a conserved elongation factor and transcription regulation in budding and fission yeast. *Genome Research*, 26(6):799–811, jun 2016. ISSN 1088-9051. doi: 10.1101/gr.204578.116. URL <http://genome.cshlp.org/lookup/doi/10.1101/gr.204578.116>.

- [24] G. T. Booth, P. K. Parua, M. Sansó, R. P. Fisher, and J. T. Lis. Cdk9 regulates a promoter-proximal checkpoint to modulate RNA polymerase II elongation rate in fission yeast. *Nature Communications*, 9(1):543, dec 2018. ISSN 2041-1723. doi: 10.1038/s41467-018-03006-4. URL <http://www.nature.com/articles/s41467-018-03006-4>.
- [25] M. J. Booth, T. W. B. Ost, D. Beraldi, N. M. Bell, M. R. Branco, W. Reik, and S. Balasubramanian. Oxidative bisulfite sequencing of 5-methylcytosine and 5-hydroxymethylcytosine. *Nature Protocols*, 8(10):1841–1851, oct 2013. ISSN 1754-2189. doi: 10.1038/nprot.2013.115. URL <http://www.nature.com/articles/nprot.2013.115>.
- [26] D. Bourc’his and T. H. Bestor. Meiotic catastrophe and retrotransposon reactivation in male germ cells lacking Dnmt3L. *Nature*, 431(7004):96–99, sep 2004. ISSN 0028-0836. doi: 10.1038/nature02886. URL <http://www.nature.com/articles/nature02886>.
- [27] P. Brazauskas. *The Biological Role of DNA modifications deposited over gene bodies*. Dphil, Oxford University, 2015.
- [28] E. Calo and J. Wysocka. Modification of Enhancer Chromatin: What, How, and Why? *Molecular Cell*, 49(5):825–837, mar 2013. ISSN 10972765. doi: 10.1016/j.molcel.2013.01.038. URL <https://linkinghub.elsevier.com/retrieve/pii/S1097276513001020>.
- [29] J. Cannuyer, A. Van Tongelen, A. Lorient, and C. De Smet. A gene expression signature identifying transient DNMT1 depletion as a causal factor of cancer-germline gene activation in melanoma. *Clinical Epigenetics*, 7(1):114, dec 2015. ISSN 1868-7075. doi: 10.1186/s13148-015-0147-4. URL <https://clinicalepigeneticsjournal.biomedcentral.com/articles/10.1186/s13148-015-0147-4>.
- [30] M. Castelnovo, J. B. Zaugg, E. Guffanti, A. Maffioletti, J. Camblong, Z. Xu, S. Clauder-Münster, L. M. Steinmetz, N. M. Luscombe, and F. Stutz. Role

- of histone modifications and early termination in pervasive transcription and antisense-mediated gene silencing in yeast. *Nucleic Acids Research*, 42(7):4348–4362, apr 2014. ISSN 1362-4962. doi: 10.1093/nar/gku100. URL <https://academic.oup.com/nar/article-lookup/doi/10.1093/nar/gku100>.
- [31] S. Catania, P. A. Dumesic, H. Pimentel, A. Nasif, C. I. Stoddard, J. E. Burke, J. K. Diedrich, S. Cooke, T. Shea, E. Gienger, R. Lintner, J. R. Yates, P. Hajkova, G. J. Narlikar, C. A. Cuomo, J. K. Pritchard, and H. D. Madhani. Evolutionary Persistence of DNA Methylation for Millions of Years after Ancient Loss of a De Novo Methyltransferase. *Cell*, 180(2):263–277.e20, jan 2020. ISSN 00928674. doi: 10.1016/j.cell.2019.12.012. URL <https://linkinghub.elsevier.com/retrieve/pii/S0092867419313741>.
- [32] R. G. Cavalcante and M. A. Sartor. annotatr: genomic regions in context. *Bioinformatics*, 33(15):2381–2383, aug 2017. ISSN 1367-4803. doi: 10.1093/bioinformatics/btx183. URL <https://academic.oup.com/bioinformatics/article/33/15/2381/3092365>.
- [33] H.-Y. Chen, W.-L. Zhang, L. Zhang, P. Yang, F. Li, Z.-R. Yang, J. Wang, M. Pang, Y. Hong, C. Yan, W. Li, J. Liu, N. Xu, L. Chen, X.-B. Xiao, Y. Qin, X.-H. He, H. Liu, H.-C. Zhu, C. He, J. Lin, and H.-M. Jing. 5-Hydroxymethylcytosine profiles of cfDNA are highly predictive of R-CHOP treatment response in diffuse large B cell lymphoma patients. *Clinical Epigenetics*, 13(1):33, dec 2021. ISSN 1868-7075. doi: 10.1186/s13148-020-00973-8. URL <https://clinicalepigeneticsjournal.biomedcentral.com/articles/10.1186/s13148-020-00973-8>.
- [34] R. Chhabra. miRNA and Methylation: A Multifaceted Liaison. *ChemBioChem*, 16(2):195–203, jan 2015. ISSN 14394227. doi: 10.1002/cbic.201402449. URL <https://onlinelibrary.wiley.com/doi/10.1002/cbic.201402449>.
- [35] J. S. Choy, S. Wei, J. Y. Lee, S. Tan, S. Chu, and T.-H. Lee. DNA Methylation Increases Nucleosome Compaction and Rigidity. *Journal of the American Chemi-*

- cal Society*, 132(6):1782–1783, feb 2010. ISSN 0002-7863. doi: 10.1021/ja910264z. URL <https://pubs.acs.org/doi/10.1021/ja910264z>.
- [36] J. K. Christman. 5-Azacytidine and 5-aza-2-deoxycytidine as inhibitors of DNA methylation: mechanistic studies and their implications for cancer therapy. *Oncogene*, 21(35):5483–5495, aug 2002. ISSN 0950-9232. doi: 10.1038/sj.onc.1205699. URL <http://www.nature.com/articles/1205699>.
- [37] C. K. Collings, P. J. Waddell, and J. N. Anderson. Effects of DNA methylation on nucleosome stability. *Nucleic Acids Research*, 41(5):2918–2931, mar 2013. ISSN 1362-4962. doi: 10.1093/nar/gks893. URL <https://academic.oup.com/nar/article-lookup/doi/10.1093/nar/gks893>.
- [38] L. Core and K. Adelman. Promoter-proximal pausing of RNA polymerase II: a nexus of gene regulation. *Genes & Development*, 33(15-16):960–982, aug 2019. ISSN 0890-9369. doi: 10.1101/gad.325142.119. URL <http://genesdev.cshlp.org/lookup/doi/10.1101/gad.325142.119>.
- [39] L. J. Core, J. J. Waterfall, and J. T. Lis. Nascent RNA sequencing reveals widespread pausing and divergent initiation at human promoters. *Science*, 2008. ISSN 00368075. doi: 10.1126/science.1162228.
- [40] L. J. Core, A. L. Martins, C. G. Danko, C. T. Waters, A. Siepel, and J. T. Lis. Analysis of nascent RNA identifies a unified architecture of initiation regions at mammalian promoters and enhancers. *Nature Genetics*, 46(12):1311–1320, dec 2014. ISSN 1061-4036. doi: 10.1038/ng.3142. URL <http://www.nature.com/articles/ng.3142>.
- [41] D. Cortázar, C. Kunz, J. Selfridge, T. Lettieri, Y. Saito, E. MacDougall, A. Wirz, D. Schuermann, A. L. Jacobs, F. Siegrist, R. Steinacher, J. Jiricny, A. Bird, and P. Schär. Embryonic lethal phenotype reveals a function of TDG in maintaining epigenetic stability. *Nature*, 470(7334):419–423, feb 2011. ISSN 0028-0836. doi: 10.1038/nature09672. URL <http://www.nature.com/articles/nature09672>.

- [42] S. Cortellino, J. Xu, M. Sannai, R. Moore, E. Caretti, A. Cigliano, M. Le Coz, K. Devarajan, A. Wessels, D. Soprano, L. K. Abramowitz, M. S. Bartolomei, F. Rambow, M. R. Bassi, T. Bruno, M. Fanciulli, C. Renner, A. J. Klein-Szanto, Y. Matsumoto, D. Kobi, I. Davidson, C. Alberti, L. Larue, and A. Bellacosa. Thymine DNA Glycosylase Is Essential for Active DNA Demethylation by Linked Deamination-Base Excision Repair. *Cell*, 146(1):67–79, jul 2011. ISSN 00928674. doi: 10.1016/j.cell.2011.06.020. URL <https://linkinghub.elsevier.com/retrieve/pii/S0092867411006623>.
- [43] M. P. Creyghton, A. W. Cheng, G. G. Welstead, T. Kooistra, B. W. Carey, E. J. Steine, J. Hanna, M. A. Lodato, G. M. Frampton, P. A. Sharp, L. A. Boyer, R. A. Young, and R. Jaenisch. Histone H3K27ac separates active from poised enhancers and predicts developmental state. *Proceedings of the National Academy of Sciences*, 107(50):21931–21936, dec 2010. ISSN 0027-8424. doi: 10.1073/pnas.1016071107. URL <http://www.pnas.org/cgi/doi/10.1073/pnas.1016071107>.
- [44] X.-L. Cui, J. Nie, J. Ku, U. Dougherty, D. C. West-Szymanski, F. Collin, C. K. Ellison, L. Sieh, Y. Ning, Z. Deng, C. W. T. Zhao, A. Bergamaschi, J. Pekow, J. Wei, A. V. Beadell, Z. Zhang, G. Sharma, R. Talwar, P. Arensdorf, J. Karpus, A. Goel, M. Bissonnette, W. Zhang, S. Levy, and C. He. A human tissue map of 5-hydroxymethylcytosines exhibits tissue specificity through gene and enhancer modulation. *Nature Communications*, 11(1):6161, dec 2020. ISSN 2041-1723. doi: 10.1038/s41467-020-20001-w. URL <http://www.nature.com/articles/s41467-020-20001-w>.
- [45] S. R. Dalton and A. Bellacosa. DNA demethylation by TDG. *Epigenomics*, 4(4):459–467, aug 2012. ISSN 1750-1911. doi: 10.2217/epi.12.36. URL <https://www.futuremedicine.com/doi/10.2217/epi.12.36>.
- [46] C. G. Danko, N. Hah, X. Luo, A. L. Martins, L. Core, J. T. Lis, A. Siepel, and W. L. Kraus. Signaling Pathways Differentially Affect RNA Polymerase II Initiation, Pausing, and Elongation Rate in Cells. *Molecular Cell*, 50(2):212–

- 222, apr 2013. ISSN 10972765. doi: 10.1016/j.molcel.2013.02.015. URL <https://linkinghub.elsevier.com/retrieve/pii/S1097276513001718>.
- [47] J. E. DeNizio, B. J. Dow, J. C. Serrano, U. Ghanty, A. C. Drohat, and R. M. Kohli. TET-TDG Active DNA Demethylation at CpG and Non-CpG Sites. *Journal of Molecular Biology*, 433(8):166877, apr 2021. ISSN 00222836. doi: 10.1016/j.jmb.2021.166877. URL <https://linkinghub.elsevier.com/retrieve/pii/S0022283621000711>.
- [48] M. Depken, J. M. Parrondo, and S. W. Grill. Intermittent Transcription Dynamics for the Rapid Production of Long Transcripts of High Fidelity. *Cell Reports*, 5(2):521–530, oct 2013. ISSN 22111247. doi: 10.1016/j.celrep.2013.09.007. URL <https://linkinghub.elsevier.com/retrieve/pii/S2211124713005159>.
- [49] S. Deshmukh, V. C. Ponnaluri, N. Dai, S. Pradhan, and D. Deobagkar. Levels of DNA cytosine methylation in the Drosophila genome. *PeerJ*, 6(7):e5119, jul 2018. ISSN 2167-8359. doi: 10.7717/peerj.5119. URL <https://peerj.com/articles/5119>.
- [50] A. Dobin, C. A. Davis, F. Schlesinger, J. Drenkow, C. Zaleski, S. Jha, P. Batut, M. Chaisson, and T. R. Gingeras. STAR: ultrafast universal RNA-seq aligner. *Bioinformatics*, 29(1):15–21, jan 2013. ISSN 1460-2059. doi: 10.1093/bioinformatics/bts635. URL <https://academic.oup.com/bioinformatics/article-lookup/doi/10.1093/bioinformatics/bts635>.
- [51] K. M. Dorigi, T. Swigut, T. Henriques, N. V. Bhanu, B. S. Scruggs, N. Nady, C. D. Still, B. A. Garcia, K. Adelman, and J. Wysocka. Mll3 and Mll4 Facilitate Enhancer RNA Synthesis and Transcription from Promoters Independently of H3K4 Monomethylation. *Molecular Cell*, 66(4):568–576.e4, may 2017. ISSN 10972765. doi: 10.1016/j.molcel.2017.04.018. URL <https://linkinghub.elsevier.com/retrieve/pii/S1097276517302745>.
- [52] N. Dukler, G. T. Booth, Y.-F. Huang, N. Tippens, C. T. Waters, C. G. Danko, J. T. Lis, and A. Siepel. Nascent RNA sequencing reveals a dynamic global tran-

- scriptional response at genes and enhancers to the natural medicinal compound celastrol. *Genome Research*, 27(11):1816–1829, nov 2017. ISSN 1088-9051. doi: 10.1101/gr.222935.117. URL <http://genome.cshlp.org/lookup/doi/10.1101/gr.222935.117>.
- [53] S. H. Duttke, S. A. Lacadie, M. M. Ibrahim, C. K. Glass, D. L. Corcoran, C. Benner, S. Heinz, J. T. Kadonaga, and U. Ohler. Human Promoters Are Intrinsically Directional. *Molecular Cell*, 57(4):674–684, feb 2015. ISSN 10972765. doi: 10.1016/j.molcel.2014.12.029. URL <https://linkinghub.elsevier.com/retrieve/pii/S1097276514010077>.
- [54] S. Ecsedi, J. Rodríguez-Aguilera, and H. Hernandez-Vargas. 5-Hydroxymethylcytosine (5hmC), or How to Identify Your Favorite Cell. *Epigenomes*, 2(1):3, jan 2018. ISSN 2075-4655. doi: 10.3390/epigenomes2010003. URL <http://www.mdpi.com/2075-4655/2/1/3>.
- [55] A. H. Ehrensberger, G. P. Kelly, and J. Q. Svejstrup. Mechanistic Interpretation of Promoter-Proximal Peaks and RNAPII Density Maps. *Cell*, 154(4):713–715, aug 2013. ISSN 00928674. doi: 10.1016/j.cell.2013.07.032. URL <https://linkinghub.elsevier.com/retrieve/pii/S0092867413009033>.
- [56] M. Ehrlich. DNA hypomethylation in cancer cells. *Epigenomics*, 1(2):239–259, dec 2009. ISSN 1750-1911. doi: 10.2217/epi.09.33. URL <https://www.futuremedicine.com/doi/10.2217/epi.09.33>.
- [57] M. Ehrlich, M. A. Gama-Sosa, L.-H. Huang, R. M. Midgett, K. C. Kuo, R. A. McCune, and C. Gehrke. Amount and distribution of 5-methylcytosine in human DNA from different types of tissues or cells. *Nucleic Acids Research*, 10(8):2709–2721, 1982. ISSN 0305-1048. doi: 10.1093/nar/10.8.2709. URL <https://academic.oup.com/nar/article-lookup/doi/10.1093/nar/10.8.2709>.
- [58] A. R. Elhamamsy. Role of DNA methylation in imprinting disorders: an updated review. *Journal of Assisted Reproduction and Genetics*, 34(5):549–562, may 2017.

ISSN 1058-0468. doi: 10.1007/s10815-017-0895-5. URL <http://link.springer.com/10.1007/s10815-017-0895-5>.

- [59] E. N. Elliott, K. L. Sheaffer, and K. H. Kaestner. The ‘de novo’ DNA methyltransferase Dnmt3b compensates the Dnmt1-deficient intestinal epithelium. *eLife*, 5, jan 2016. ISSN 2050-084X. doi: 10.7554/eLife.12975. URL <https://elifesciences.org/articles/12975>.
- [60] N. D. Elrod, T. Henriques, K.-L. Huang, D. C. Tatomer, J. E. Wilusz, E. J. Wagner, and K. Adelman. The Integrator Complex Attenuates Promoter-Proximal Transcription at Protein-Coding Genes. *Molecular Cell*, 76(5):738–752.e7, dec 2019. ISSN 10972765. doi: 10.1016/j.molcel.2019.10.034. URL <https://www.cambridge.org/core/product/identifier/CB09781107415324A009/type/book{ }parthttp://arxiv.org/abs/1011.1669http://dx.doi.org/10.1088/1751-8113/44/8/085201https://iopscience.iop.org/article/10.1088/1751-8113/44/8/085201https://linkinghub.elsevi>.
- [61] C. Engel, T. Gubbey, S. Neyer, S. Sainsbury, C. Oberthuer, C. Baejen, C. Bernecky, and P. Cramer. Structural Basis of RNA Polymerase I Transcription Initiation. *Cell*, 169(1):120–131.e22, mar 2017. ISSN 00928674. doi: 10.1016/j.cell.2017.03.003. URL <https://linkinghub.elsevier.com/retrieve/pii/S0092867417302519>.
- [62] P. Ewels, M. Magnusson, S. Lundin, and M. Käller. MultiQC: summarize analysis results for multiple tools and samples in a single report. *Bioinformatics*, 32(19):3047–3048, oct 2016. ISSN 1367-4803. doi: 10.1093/bioinformatics/btw354. URL <https://academic.oup.com/bioinformatics/article-lookup/doi/10.1093/bioinformatics/btw354>.
- [63] G. Fan, C. Beard, R. Z. Chen, G. Csankovszki, Y. Sun, M. Siniaia, D. Biniszkievicz, B. Bates, P. P. Lee, R. Kühn, A. Trumpp, C.-S. Poon, C. B. Wilson, and R. Jaenisch. DNA Hypomethylation Perturbs the Function and Survival of CNS Neurons in Postnatal Animals. *The Journal of Neuroscience*, 21(3):

- 788–797, feb 2001. ISSN 0270-6474. doi: 10.1523/JNEUROSCI.21-03-00788.2001. URL <https://www.jneurosci.org/lookup/doi/10.1523/JNEUROSCI.21-03-00788.2001>.
- [64] J. Feng, Y. Zhou, S. L. Campbell, T. Le, E. Li, J. D. Sweatt, A. J. Silva, and G. Fan. Dnmt1 and Dnmt3a maintain DNA methylation and regulate synaptic function in adult forebrain neurons. *Nature Neuroscience*, 13(4):423–430, apr 2010. ISSN 1097-6256. doi: 10.1038/nn.2514. URL <http://www.nature.com/articles/nn.2514>.
- [65] I. S. Fetahu, D. Ma, K. Rabidou, C. Argueta, M. Smith, H. Liu, F. Wu, and Y. G. Shi. Epigenetic signatures of methylated DNA cytosine in Alzheimer’s disease. *Science Advances*, 5(8):eaaw2880, aug 2019. ISSN 2375-2548. doi: 10.1126/sciadv.aaw2880. URL <https://advances.sciencemag.org/lookup/doi/10.1126/sciadv.aaw2880>.
- [66] G. Ficiz, M. R. Branco, S. Seisenberger, F. Santos, F. Krueger, T. A. Hore, C. J. Marques, S. Andrews, and W. Reik. Dynamic regulation of 5-hydroxymethylcytosine in mouse ES cells and during differentiation. *Nature*, 473(7347):398–402, may 2011. ISSN 0028-0836. doi: 10.1038/nature10008. URL <http://www.nature.com/articles/nature10008>.
- [67] M. E. Figueroa, O. Abdel-Wahab, C. Lu, P. S. Ward, J. Patel, A. Shih, Y. Li, N. Bhagwat, A. Vasanthakumar, H. F. Fernandez, M. S. Tallman, Z. Sun, K. Wolniak, J. K. Peeters, W. Liu, S. E. Choe, V. R. Fantin, E. Paietta, B. Löwenberg, J. D. Licht, L. A. Godley, R. Delwel, P. J. Valk, C. B. Thompson, R. L. Levine, and A. Melnick. Leukemic IDH1 and IDH2 Mutations Result in a Hypermethylation Phenotype, Disrupt TET2 Function, and Impair Hematopoietic Differentiation. *Cancer Cell*, 18(6):553–567, dec 2010. ISSN 15356108. doi: 10.1016/j.ccr.2010.11.015. URL <https://linkinghub.elsevier.com/retrieve/pii/S1535610810004836>.
- [68] W. J. Fitzsimmons, R. J. Woods, J. T. McCrone, A. Woodman, J. J. Arnold,

- M. Yennawar, R. Evans, C. E. Cameron, and A. S. Luring. A speed–fidelity trade-off determines the mutation rate and virulence of an RNA virus. *PLOS Biology*, 16(6):e2006459, jun 2018. ISSN 1545-7885. doi: 10.1371/journal.pbio.2006459. URL <https://dx.plos.org/10.1371/journal.pbio.2006459>.
- [69] N. Fong, K. Brannan, B. Erickson, H. Kim, M. A. Cortazar, R. M. Sheridan, T. Nguyen, S. Karp, and D. L. Bentley. Effects of Transcription Elongation Rate and Xrn2 Exonuclease Activity on RNA Polymerase II Termination Suggest Widespread Kinetic Competition. *Molecular Cell*, 2015. ISSN 10974164. doi: 10.1016/j.molcel.2015.09.026.
- [70] N. H. Freese, D. C. Norris, and A. E. Loraine. Integrated genome browser: visual analytics platform for genomics. *Bioinformatics*, 32(14): 2089–2095, jul 2016. ISSN 1367-4803. doi: 10.1093/bioinformatics/btw069. URL <https://academic.oup.com/bioinformatics/article-lookup/doi/10.1093/bioinformatics/btw069>.
- [71] R. Fujiwara, N. Damodaren, J. E. Wilusz, and K. Murakami. The capping enzyme facilitates promoter escape and assembly of a follow-on preinitiation complex for reinitiation. *Proceedings of the National Academy of Sciences*, 116(45):22573–22582, nov 2019. ISSN 0027-8424. doi: 10.1073/pnas.1905449116. URL <http://www.pnas.org/lookup/doi/10.1073/pnas.1905449116>.
- [72] T. Fukaya, B. Lim, and M. Levine. Rapid Rates of Pol II Elongation in the Drosophila Embryo. *Current Biology*, 27(9):1387–1391, may 2017. ISSN 09609822. doi: 10.1016/j.cub.2017.03.069. URL <https://linkinghub.elsevier.com/retrieve/pii/S0960982217303998>.
- [73] T. Gao and J. Qian. EnhancerAtlas 2.0: an updated resource with enhancer annotation in 586 tissue/cell types across nine species. *Nucleic Acids Research*, nov 2019. ISSN 0305-1048. doi: 10.1093/nar/gkz980. URL <https://academic.oup.com/nar/advance-article/doi/10.1093/nar/gkz980/5628925>.
- [74] M. Gardiner-Garden and M. Frommer. CpG Islands in vertebrate genomes. *Jour-*

- nal of Molecular Biology*, 196(2):261–282, jul 1987. ISSN 00222836. doi: 10.1016/0022-2836(87)90689-9. URL <https://linkinghub.elsevier.com/retrieve/pii/0022283687906899>.
- [75] F. Gaudet. Induction of Tumors in Mice by Genomic Hypomethylation. *Science*, 300(5618):489–492, apr 2003. ISSN 00368075. doi: 10.1126/science.1083558. URL <https://www.sciencemag.org/lookup/doi/10.1126/science.1083558>.
- [76] B. Gelfand, J. Mead, A. Bruning, N. Apostolopoulos, V. Tadigotla, V. Nagaraj, A. M. Sengupta, and A. K. Vershon. Regulated Antisense Transcription Controls Expression of Cell-Type-Specific Genes in Yeast. *Molecular and Cellular Biology*, 31(8):1701–1709, apr 2011. ISSN 0270-7306. doi: 10.1128/MCB.01071-10. URL <http://mcb.asm.org/cgi/doi/10.1128/MCB.01071-10>.
- [77] D. Globisch, M. Münzel, M. Müller, S. Michalakakis, M. Wagner, S. Koch, T. Brückl, M. Biel, and T. Carell. Tissue Distribution of 5-Hydroxymethylcytosine and Search for Active Demethylation Intermediates. *PLoS ONE*, 5(12):e15367, dec 2010. ISSN 1932-6203. doi: 10.1371/journal.pone.0015367. URL <https://dx.plos.org/10.1371/journal.pone.0015367>.
- [78] C. R. Good, J. Madzo, B. Patel, S. Maegawa, N. Engel, J. Jelinek, and J.-P. J. Issa. A novel isoform of TET1 that lacks a CXXC domain is overexpressed in cancer. *Nucleic Acids Research*, 45(14):8269–8281, aug 2017. ISSN 0305-1048. doi: 10.1093/nar/gkx435. URL <http://academic.oup.com/nar/article/45/14/8269/3844757>.
- [79] J. R. Graff, J. G. Herman, S. Myöhänen, S. B. Baylin, and P. M. Vertino. Mapping Patterns of CpG Island Methylation in Normal and Neoplastic Cells Implicates Both Upstream and Downstream Regions in de Novo Methylation. *Journal of Biological Chemistry*, 272(35):22322–22329, aug 1997. ISSN 00219258. doi: 10.1074/jbc.272.35.22322. URL <https://linkinghub.elsevier.com/retrieve/pii/S002192581965698X>.
- [80] S. Gressel, B. Schwalb, T. M. Decker, W. Qin, H. Leonhardt, D. Eick, and

- P. Cramer. CDK9-dependent RNA polymerase II pausing controls transcription initiation. *eLife*, 6, 2017. ISSN 2050-084X. doi: 10.7554/eLife.29736.
- [81] P. Grossmann, A. Lück, and C. Kaleta. Model-based genome-wide determination of RNA chain elongation rates in *Escherichia coli*. *Scientific Reports*, 7(1):17213, dec 2017. ISSN 2045-2322. doi: 10.1038/s41598-017-17408-9. URL <http://www.nature.com/articles/s41598-017-17408-9>.
- [82] V. Haberle and A. Stark. Eukaryotic core promoters and the functional basis of transcription initiation. *Nature Reviews Molecular Cell Biology*, 19(10):621–637, oct 2018. ISSN 1471-0072. doi: 10.1038/s41580-018-0028-8. URL <http://www.nature.com/articles/s41580-018-0028-8>.
- [83] M. A. Hahn, R. Qiu, X. Wu, A. X. Li, H. Zhang, J. Wang, J. Jui, S.-G. Jin, Y. Jiang, G. P. Pfeifer, and Q. Lu. Dynamics of 5-Hydroxymethylcytosine and Chromatin Marks in Mammalian Neurogenesis. *Cell Reports*, 3(2):291–300, feb 2013. ISSN 22111247. doi: 10.1016/j.celrep.2013.01.011. URL <https://linkinghub.elsevier.com/retrieve/pii/S2211124713000181>.
- [84] M. A. Hahn, P. E. Szabó, and G. P. Pfeifer. 5-Hydroxymethylcytosine: A stable or transient DNA modification? *Genomics*, 104(5):314–323, nov 2014. ISSN 08887543. doi: 10.1016/j.ygeno.2014.08.015. URL <https://linkinghub.elsevier.com/retrieve/pii/S0888754314001578>.
- [85] A. Harrison and A. Parle-McDermott. DNA Methylation: A Timeline of Methods and Applications. *Frontiers in Genetics*, 2, 2011. ISSN 1664-8021. doi: 10.3389/fgene.2011.00074. URL <http://journal.frontiersin.org/article/10.3389/fgene.2011.00074/abstract>.
- [86] K. Hata and Y. Sakaki. Identification of critical CpG sites for repression of L1 transcription by DNA methylation. *Gene*, 189(2):227–234, apr 1997. ISSN 03781119. doi: 10.1016/S0378-1119(96)00856-6. URL <https://linkinghub.elsevier.com/retrieve/pii/S0378111996008566>.

- [87] H. Hayatsu and M. Shiragami. Reaction of bisulfite with the 5-hydroxymethyl group in pyrimidines and in phage DNAs. *Biochemistry*, 18(4):632–637, feb 1979. ISSN 0006-2960. doi: 10.1021/bi00571a013. URL <https://pubs.acs.org/doi/abs/10.1021/bi00571a013>.
- [88] H. Hayatsu, Y. Wataya, and K. Kai. Addition of sodium bisulfite to uracil and to cytosine. *Journal of the American Chemical Society*, 92(3):724–726, feb 1970. ISSN 0002-7863. doi: 10.1021/ja00706a062. URL <https://pubs.acs.org/doi/abs/10.1021/ja00706a062>.
- [89] H. Hayatsu, Y. Wataya, K. Kai, and S. Iida. Reaction of sodium bisulfite with uracil, cytosine, and their derivatives. *Biochemistry*, 9(14):2858–2865, jul 1970. ISSN 0006-2960. doi: 10.1021/bi00816a016. URL <https://pubs.acs.org/doi/abs/10.1021/bi00816a016>.
- [90] Y.-F. He, B.-Z. Li, Z. Li, P. Liu, Y. Wang, Q. Tang, J. Ding, Y. Jia, Z. Chen, L. Li, Y. Sun, X. Li, Q. Dai, C.-X. Song, K. Zhang, C. He, and G.-L. Xu. Tet-Mediated Formation of 5-Carboxylcytosine and Its Excision by TDG in Mammalian DNA. *Science*, 333(6047):1303–1307, sep 2011. ISSN 0036-8075. doi: 10.1126/science.1210944. URL <https://www.sciencemag.org/lookup/doi/10.1126/science.1210944>.
- [91] S. Heinz, C. Benner, N. Spann, E. Bertolino, Y. C. Lin, P. Laslo, J. X. Cheng, C. Murre, H. Singh, and C. K. Glass. Simple Combinations of Lineage-Determining Transcription Factors Prime cis-Regulatory Elements Required for Macrophage and B Cell Identities. *Molecular Cell*, 38(4):576–589, may 2010. ISSN 10972765. doi: 10.1016/j.molcel.2010.05.004. URL <https://linkinghub.elsevier.com/retrieve/pii/S1097276510003667>.
- [92] I. R. Henderson and S. E. Jacobsen. Epigenetic inheritance in plants. *Nature*, 447(7143):418–424, may 2007. ISSN 0028-0836. doi: 10.1038/nature05917. URL <http://www.nature.com/articles/nature05917>.
- [93] T. Henriques, B. S. Scruggs, M. O. Inouye, G. W. Muse, L. H. Williams, A. B.

- Burkholder, C. A. Lavender, D. C. Fargo, and K. Adelman. Widespread transcriptional pausing and elongation control at enhancers. *Genes & Development*, 32(1):26–41, jan 2018. ISSN 0890-9369. doi: 10.1101/gad.309351.117. URL <http://genesdev.cshlp.org/lookup/doi/10.1101/gad.309351.117>.
- [94] J. Hetzel, S. H. Duttke, C. Benner, and J. Chory. Nascent RNA sequencing reveals distinct features in plant transcription. *Proceedings of the National Academy of Sciences*, 113(43):12316–12321, oct 2016. ISSN 0027-8424. doi: 10.1073/pnas.1603217113. URL <http://www.pnas.org/lookup/doi/10.1073/pnas.1603217113>.
- [95] E. Hodges, A. Molaro, C. O. Dos Santos, P. Thekkat, Q. Song, P. J. Uren, J. Park, J. Butler, S. Rafii, W. R. McCombie, A. D. Smith, and G. J. Hannon. Directional DNA Methylation Changes and Complex Intermediate States Accompany Lineage Specificity in the Adult Hematopoietic Compartment. *Molecular Cell*, 44(1):17–28, oct 2011. ISSN 10972765. doi: 10.1016/j.molcel.2011.08.026. URL <https://linkinghub.elsevier.com/retrieve/pii/S1097276511006769>.
- [96] R. Holliday and J. Pugh. DNA modification mechanisms and gene activity during development. *Science*, 187(4173):226–232, jan 1975. ISSN 0036-8075. doi: 10.1126/science.1111098. URL <https://www.sciencemag.org/lookup/doi/10.1126/science.1111098>.
- [97] C. F. Hongay, P. L. Grisafi, T. Galitski, and G. R. Fink. Antisense Transcription Controls Cell Fate in *Saccharomyces cerevisiae*. *Cell*, 127(4):735–745, nov 2006. ISSN 00928674. doi: 10.1016/j.cell.2006.09.038. URL <https://linkinghub.elsevier.com/retrieve/pii/S0092867406013006>.
- [98] L. Hou, X. Wang, P. Yang, B. Li, Z. Lin, L. Kang, and X. Wang. DNA methyltransferase 3 participates in behavioral phase change in the migratory locust. *Insect Biochemistry and Molecular Biology*, 121:103374, jun 2020. ISSN 09651748. doi: 10.1016/j.ibmb.2020.103374. URL <https://linkinghub.elsevier.com/retrieve/pii/S0965174820300631>.

- [99] L. Hu, J. Lu, J. Cheng, Q. Rao, Z. Li, H. Hou, Z. Lou, L. Zhang, W. Li, W. Gong, M. Liu, C. Sun, X. Yin, J. Li, X. Tan, P. Wang, Y. Wang, D. Fang, Q. Cui, P. Yang, C. He, H. Jiang, C. Luo, and Y. Xu. Structural insight into substrate preference for TET-mediated oxidation. *Nature*, 527(7576):118–122, nov 2015. ISSN 0028-0836. doi: 10.1038/nature15713. URL <http://www.nature.com/articles/nature15713>.
- [100] F. Huber, D. Bunina, I. Gupta, A. Khmelinskii, M. Meurer, P. Theer, L. M. Steinmetz, and M. Knop. Protein Abundance Control by Non-coding Antisense Transcription. *Cell Reports*, 15(12):2625–2636, jun 2016. ISSN 22111247. doi: 10.1016/j.celrep.2016.05.043. URL <https://linkinghub.elsevier.com/retrieve/pii/S221112471630626X>.
- [101] B. Institute. Picard Tools. URL <https://broadinstitute.github.io/picard/>.
- [102] J.-p. Issa. CpG island methylator phenotype in cancer. *Nature Reviews Cancer*, 4(12):988–993, dec 2004. ISSN 1474-175X. doi: 10.1038/nrc1507. URL <http://www.nature.com/articles/nrc1507>.
- [103] S. Ito, A. C. D’Alessio, O. V. Taranova, K. Hong, L. C. Sowers, and Y. Zhang. Role of Tet proteins in 5mC to 5hmC conversion, ES-cell self-renewal and inner cell mass specification. *Nature*, 466(7310):1129–1133, aug 2010. ISSN 0028-0836. doi: 10.1038/nature09303. URL <https://www.ncbi.nlm.nih.gov/pubmed/20639862><http://www.nature.com/articles/nature09303>.
- [104] S. Ito, L. Shen, Q. Dai, S. C. Wu, L. B. Collins, J. A. Swenberg, C. He, and Y. Zhang. Tet Proteins Can Convert 5-Methylcytosine to 5-Formylcytosine and 5-Carboxylcytosine. *Science*, 333(6047):1300–1303, sep 2011. ISSN 0036-8075. doi: 10.1126/science.1210597. URL <https://www.sciencemag.org/lookup/doi/10.1126/science.1210597>.
- [105] M. Jain, S. Koren, K. H. Miga, J. Quick, A. C. Rand, T. A. Sasani, J. R. Tyson, A. D. Beggs, A. T. Dilthey, I. T. Fiddes, S. Malla, H. Marriott, T. Nieto, J. O’Grady, H. E. Olsen, B. S. Pedersen, A. Rhie, H. Richardson, A. R. Quinlan,

- T. P. Snutch, L. Tee, B. Paten, A. M. Phillippy, J. T. Simpson, N. J. Loman, and M. Loose. Nanopore sequencing and assembly of a human genome with ultra-long reads. *Nature Biotechnology*, 36(4):338–345, apr 2018. ISSN 1087-0156. doi: 10.1038/nbt.4060. URL <http://www.nature.com/articles/nbt.4060>.
- [106] M. Janoušková, Z. Vaníková, F. Nici, S. Boháčová, D. Vítovská, H. Šanderová, M. Hocek, and L. Krásný. 5-(Hydroxymethyl)uracil and -cytosine as potential epigenetic marks enhancing or inhibiting transcription with bacterial RNA polymerase. *Chemical Communications*, 53(99):13253–13255, 2017. ISSN 1359-7345. doi: 10.1039/C7CC08053K. URL <http://xlink.rsc.org/?DOI=C7CC08053K>.
- [107] C. Jin, Y. Lu, J. Jelinek, S. Liang, M. R. H. Estecio, M. C. Barton, and J.-P. J. Issa. TET1 is a maintenance DNA demethylase that prevents methylation spreading in differentiated cells. *Nucleic Acids Research*, 42(11):6956–6971, jun 2014. ISSN 0305-1048. doi: 10.1093/nar/gku372. URL <https://www.ncbi.nlm.nih.gov/pubmed/24875481https://academic.oup.com/nar/article-lookup/doi/10.1093/nar/gku372>.
- [108] D. Jjingo, A. B. Conley, S. V. Yi, V. V. Lunyak, and I. K. Jordan. On the presence and role of human gene-body DNA methylation. *Oncotarget*, 3(4):462–474, apr 2012. ISSN 1949-2553. doi: 10.18632/oncotarget.497. URL <https://www.oncotarget.com/lookup/doi/10.18632/oncotarget.497>.
- [109] P. A. Jones. The DNA methylation paradox. *Trends in Genetics*, 15(1):34–37, jan 1999. ISSN 01689525. doi: 10.1016/S0168-9525(98)01636-9. URL <https://linkinghub.elsevier.com/retrieve/pii/S0168952598016369>.
- [110] P. A. Jones. Functions of DNA methylation: islands, start sites, gene bodies and beyond. *Nature Reviews Genetics*, 13(7):484–492, jul 2012. ISSN 1471-0056. doi: 10.1038/nrg3230. URL <http://www.nature.com/articles/nrg3230>.
- [111] I. Jonkers, H. Kwak, and J. T. Lis. Genome-wide dynamics of Pol II elongation and its interplay with promoter proximal pausing, chromatin, and exons. *eLife*, 2014(3), 2014. ISSN 2050084X. doi: 10.7554/eLife.02407.

- [112] A. Jordán-Pla, M. E. Pérez-Martínez, and J. E. Pérez-Ortín. Measuring RNA polymerase activity genome-wide with high-resolution run-on-based methods. *Methods*, 159-160(December 2018):177–182, apr 2019. ISSN 10462023. doi: 10.1016/j.ymeth.2019.01.017. URL <https://doi.org/10.1016/j.ymeth.2019.01.017><https://linkinghub.elsevier.com/retrieve/pii/S1046202318302950>.
- [113] B. G. Jorgensen, R. M. Berent, S. E. Ha, K. Horiguchi, K. C. Sasse, L. S. Becker, and S. Ro. DNA methylation, through DNMT1, has an essential role in the development of gastrointestinal smooth muscle cells and disease. *Cell Death & Disease*, 9(5):474, may 2018. ISSN 2041-4889. doi: 10.1038/s41419-018-0495-z. URL <http://www.nature.com/articles/s41419-018-0495-z>.
- [114] G. R. Kafer, X. Li, T. Horii, I. Suetake, S. Tajima, I. Hatada, and P. M. Carlton. 5-Hydroxymethylcytosine Marks Sites of DNA Damage and Promotes Genome Stability. *Cell Reports*, 2016. ISSN 22111247. doi: 10.1016/j.celrep.2016.01.035.
- [115] T. Kaneko-Ishino and F. Ishino. Retrotransposon silencing by DNA methylation contributed to the evolution of placentation and genomic imprinting in mammals. *Development, Growth & Differentiation*, 52(6):533–543, jul 2010. ISSN 00121592. doi: 10.1111/j.1440-169X.2010.01194.x. URL <http://doi.wiley.com/10.1111/j.1440-169X.2010.01194.x>.
- [116] O. L. Kantidze and S. V. Razin. 5-hydroxymethylcytosine in DNA repair: A new player or a red herring? *Cell Cycle*, 16(16):1499–1501, aug 2017. ISSN 1538-4101. doi: 10.1080/15384101.2017.1346761. URL <https://www.tandfonline.com/doi/full/10.1080/15384101.2017.1346761>.
- [117] W. J. Kent, A. S. Zweig, G. Barber, A. S. Hinrichs, and D. Karolchik. BigWig and BigBed: enabling browsing of large distributed datasets. *Bioinformatics*, 26(17):2204–2207, sep 2010. ISSN 1460-2059. doi: 10.1093/bioinformatics/btq351. URL <https://academic.oup.com/bioinformatics/article-lookup/doi/10.1093/bioinformatics/btq351>.
- [118] T. Khare, S. Pai, K. Koncevicus, M. Pal, E. Kriukiene, Z. Liutkeviciute, M. Ir-

- imia, P. Jia, C. Ptak, M. Xia, R. Tice, M. Tochigi, S. Moréra, A. Nazarians, D. Belsham, A. H. C. Wong, B. J. Blencowe, S. C. Wang, P. Kapranov, R. Kustra, V. Labrie, S. Klimasauskas, and A. Petronis. 5-hmC in the brain is abundant in synaptic genes and shows differences at the exon-intron boundary. *Nature Structural & Molecular Biology*, 19(10):1037–1043, oct 2012. ISSN 1545-9993. doi: 10.1038/nsmb.2372. URL <http://www.nature.com/articles/nsmb.2372>.
- [119] M. Ko, J. An, and A. Rao. DNA methylation and hydroxymethylation in hematologic differentiation and transformation. *Current Opinion in Cell Biology*, 37:91–101, dec 2015. ISSN 09550674. doi: 10.1016/j.ceb.2015.10.009. URL <https://linkinghub.elsevier.com/retrieve/pii/S0955067415001301>.
- [120] H. Kobayashi, T. Sakurai, M. Imai, N. Takahashi, A. Fukuda, O. Yayoi, S. Sato, K. Nakabayashi, K. Hata, Y. Sotomaru, Y. Suzuki, and T. Kono. Contribution of Intragenic DNA Methylation in Mouse Gametic DNA Methylomes to Establish Oocyte-Specific Heritable Marks. *PLoS Genetics*, 8(1):e1002440, jan 2012. ISSN 1553-7404. doi: 10.1371/journal.pgen.1002440. URL <https://dx.plos.org/10.1371/journal.pgen.1002440>.
- [121] R. D. Kornberg. Mediator and the mechanism of transcriptional activation. *Trends in Biochemical Sciences*, 30(5):235–239, may 2005. ISSN 09680004. doi: 10.1016/j.tibs.2005.03.011. URL <https://linkinghub.elsevier.com/retrieve/pii/S0968000405000836>.
- [122] S. Kriaucionis and N. Heintz. The Nuclear DNA Base 5-Hydroxymethylcytosine Is Present in Purkinje Neurons and the Brain. *Science*, 324(5929):929–930, may 2009. ISSN 0036-8075. doi: 10.1126/science.1169786. URL <http://www.sciencemag.org/content/324/5929/929.long><https://www.sciencemag.org/lookup/doi/10.1126/science.1169786>.
- [123] F. Krueger. Trim Galore! URL http://www.bioinformatics.babraham.ac.uk/projects/trim_galore/.
- [124] R. Kucharski, J. Maleszka, S. Foret, and R. Maleszka. Nutritional Control of

- Reproductive Status in Honeybees via DNA Methylation. *Science*, 319(5871): 1827–1830, mar 2008. ISSN 0036-8075. doi: 10.1126/science.1153069. URL <https://www.sciencemag.org/lookup/doi/10.1126/science.1153069>.
- [125] J. F. Kugel and J. A. Goodrich. Finding the start site: redefining the human initiator element. *Genes & Development*, 31(1):1–2, jan 2017. ISSN 0890-9369. doi: 10.1101/gad.295980.117. URL <http://genesdev.cshlp.org/lookup/doi/10.1101/gad.295980.117>.
- [126] H. Kwak, N. J. Fuda, L. J. Core, and J. T. Lis. Precise maps of RNA polymerase reveal how promoters direct initiation and pausing. *Science*, 339(6122): 950–953, 2013. doi: 10.1126/science.1229386. URL <https://www.ncbi.nlm.nih.gov/pubmed/23430654>.
- [127] A. J. Larsson, P. Johnsson, M. Hagemann-Jensen, L. Hartmanis, O. R. Faridani, B. Reinius, Å. Segerstolpe, C. M. Rivera, B. Ren, and R. Sandberg. Genomic encoding of transcriptional burst kinetics, 2019. ISSN 14764687.
- [128] J. A. Law and S. E. Jacobsen. Establishing, maintaining and modifying DNA methylation patterns in plants and animals. *Nature Reviews Genetics*, 11(3): 204–220, mar 2010. ISSN 1471-0056. doi: 10.1038/nrg2719. URL <http://www.nature.com/articles/nrg2719>.
- [129] K.-H. Lee, S. Oghamian, J.-A. Park, L. Kang, and P. W. Laird. The REMOTE-control system: a system for reversible and tunable control of endogenous gene expression in mice. *Nucleic Acids Research*, 45(21):12256–12269, dec 2017. ISSN 0305-1048. doi: 10.1093/nar/gkx829. URL <http://academic.oup.com/nar/article/45/21/12256/4157874>.
- [130] S. Lee, D. Cook, and M. Lawrence. plyranges: a grammar of genomic data transformation. *Genome Biology*, 20(1):4, dec 2019. ISSN 1474-760X. doi: 10.1186/s13059-018-1597-8. URL <https://genomebiology.biomedcentral.com/articles/10.1186/s13059-018-1597-8>.

- [131] T. L. Lenstra, A. Coulon, C. C. Chow, and D. R. Larson. Single-Molecule Imaging Reveals a Switch between Spurious and Functional ncRNA Transcription. *Molecular Cell*, 60(4):597–610, nov 2015. ISSN 10972765. doi: 10.1016/j.molcel.2015.09.028. URL <https://linkinghub.elsevier.com/retrieve/pii/S1097276515007674>.
- [132] C. Lepoivre, M. Belhocine, A. Bergon, A. Griffon, M. Yammine, L. Vanhille, J. Zacarias-Cabeza, M.-A. Garibal, F. Koch, M. Maqbool, R. Fenouil, B. Loriod, H. Holota, M. Gut, I. Gut, J. Imbert, J.-C. Andrau, D. Puthier, and S. Spicuglia. Divergent transcription is associated with promoters of transcriptional regulators. *BMC Genomics*, 14(1):914, 2013. ISSN 1471-2164. doi: 10.1186/1471-2164-14-914. URL <http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3882496&tool=pmcentrez&rendertype=abstract><http://bmcbgenomics.biomedcentral.com/articles/10.1186/1471-2164-14-914>.
- [133] E. Li, T. H. Bestor, and R. Jaenisch. Targeted mutation of the DNA methyltransferase gene results in embryonic lethality. *Cell*, 69(6):915–926, jun 1992. ISSN 00928674. doi: 10.1016/0092-8674(92)90611-F. URL <https://linkinghub.elsevier.com/retrieve/pii/009286749290611F>.
- [134] E. Li, C. Beard, and R. Jaenisch. Role for DNA methylation in genomic imprinting. *Nature*, 366(6453):362–365, dec 1993. ISSN 0028-0836. doi: 10.1038/366362a0. URL <http://www.nature.com/articles/366362a0>.
- [135] G. Li, X. Ruan, R. K. Auerbach, K. S. Sandhu, M. Zheng, P. Wang, H. M. Poh, Y. Goh, J. Lim, J. Zhang, H. S. Sim, S. Q. Peh, F. H. Mulawadi, C. T. Ong, Y. L. Orlov, S. Hong, Z. Zhang, S. Landt, D. Raha, G. Euskirchen, C.-L. Wei, W. Ge, H. Wang, C. Davis, K. I. Fisher-Aylor, A. Mortazavi, M. Gerstein, T. Gingeras, B. Wold, Y. Sun, M. J. Fullwood, E. Cheung, E. Liu, W.-K. Sung, M. Snyder, and Y. Ruan. Extensive Promoter-Centered Chromatin Interactions Provide a Topological Basis for Transcription Regulation. *Cell*, 148

- (1-2):84–98, jan 2012. ISSN 00928674. doi: 10.1016/j.cell.2011.12.014. URL <https://linkinghub.elsevier.com/retrieve/pii/S0092867411015170>.
- [136] H. Li, B. Handsaker, A. Wysoker, T. Fennell, J. Ruan, N. Homer, G. Marth, G. Abecasis, and R. Durbin. The Sequence Alignment/Map format and SAM-tools. *Bioinformatics*, 25(16):2078–2079, aug 2009. ISSN 1367-4803. doi: 10.1093/bioinformatics/btp352. URL <https://academic.oup.com/bioinformatics/article-lookup/doi/10.1093/bioinformatics/btp352>.
- [137] H.-D. Li, C. C. Funk, and N. D. Price. iREAD: a tool for intron retention detection from RNA-seq data. *BMC Genomics*, 21(1):128, dec 2020. ISSN 1471-2164. doi: 10.1186/s12864-020-6541-0. URL <https://bmcbgenomics.biomedcentral.com/articles/10.1186/s12864-020-6541-0>.
- [138] W. Li, X. Zhang, X. Lu, L. You, Y. Song, Z. Luo, J. Zhang, J. Nie, W. Zheng, D. Xu, Y. Wang, Y. Dong, S. Yu, J. Hong, J. Shi, H. Hao, F. Luo, L. Hua, P. Wang, X. Qian, F. Yuan, L. Wei, M. Cui, T. Zhang, Q. Liao, M. Dai, Z. Liu, G. Chen, K. Meckel, S. Adhikari, G. Jia, M. B. Bissonnette, X. Zhang, Y. Zhao, W. Zhang, C. He, and J. Liu. 5-Hydroxymethylcytosine signatures in circulating cell-free DNA as diagnostic biomarkers for human cancers. *Cell Research*, 27(10):1243–1257, oct 2017. ISSN 1001-0602. doi: 10.1038/cr.2017.121. URL <http://www.nature.com/articles/cr2017121>.
- [139] X. Li, X. Yue, W. A. Pastor, L. Lin, R. Georges, L. Chavez, S. M. Evans, and A. Rao. Tet proteins influence the balance between neuroectodermal and mesodermal fate choice by inhibiting Wnt signaling. *Proceedings of the National Academy of Sciences*, 113(51):E8267–E8276, dec 2016. ISSN 0027-8424. doi: 10.1073/pnas.1617802113. URL <http://www.pnas.org/lookup/doi/10.1073/pnas.1617802113>.
- [140] C. G. Lian, Y. Xu, C. Ceol, F. Wu, A. Larson, K. Dresser, W. Xu, L. Tan, Y. Hu, Q. Zhan, C.-w. Lee, D. Hu, B. Q. Lian, S. Kleffel, Y. Yang, J. Neiswender, A. J. Khorasani, R. Fang, C. Lezcano, L. M. Duncan, R. A. Scolyer, J. F. Thompson,

- H. Kakavand, Y. Houvras, L. I. Zon, M. C. Mihm, U. B. Kaiser, T. Schatton, B. A. Woda, G. F. Murphy, and Y. G. Shi. Loss of 5-Hydroxymethylcytosine Is an Epigenetic Hallmark of Melanoma. *Cell*, 150(6):1135–1146, sep 2012. ISSN 00928674. doi: 10.1016/j.cell.2012.07.033. URL <https://linkinghub.elsevier.com/retrieve/pii/S0092867412010124>.
- [141] Y. LIAO, J. GU, Y. WU, X. LONG, D. GE, J. XU, and J. DING. Low level of 5-Hydroxymethylcytosine predicts poor prognosis in non-small cell lung cancer. *Oncology Letters*, 11(6):3753–3760, jun 2016. ISSN 1792-1074. doi: 10.3892/ol.2016.4474. URL <https://www.spandidos-publications.com/10.3892/ol.2016.4474>.
- [142] E. Lieberman-Aiden, N. L. van Berkum, L. Williams, M. Imakaev, T. Ragoczy, A. Telling, I. Amit, B. R. Lajoie, P. J. Sabo, M. O. Dorschner, R. Sandstrom, B. Bernstein, M. A. Bender, M. Groudine, A. Gnirke, J. Stamatoyannopoulos, L. A. Mirny, E. S. Lander, and J. Dekker. Comprehensive Mapping of Long-Range Interactions Reveals Folding Principles of the Human Genome. *Science*, 326(5950):289–293, oct 2009. ISSN 0036-8075. doi: 10.1126/science.1181369. URL <https://www.science.org/doi/10.1126/science.1181369>.
- [143] B. Lim. Imaging transcriptional dynamics. *Current Opinion in Biotechnology*, 52:49–55, aug 2018. ISSN 09581669. doi: 10.1016/j.copbio.2018.02.008. URL <https://linkinghub.elsevier.com/retrieve/pii/S0958166917302367>.
- [144] I.-H. Lin, Y.-F. Chen, and M.-T. Hsu. Correlated 5-Hydroxymethylcytosine (5hmC) and Gene Expression Profiles Underpin Gene and Organ-Specific Epigenetic Regulation in Adult Mouse Brain and Liver. *PLOS ONE*, 12(1):e0170779, jan 2017. ISSN 1932-6203. doi: 10.1371/journal.pone.0170779. URL <https://dx.plos.org/10.1371/journal.pone.0170779>.
- [145] Y.-C. Lin, M. Boone, L. Meuris, I. Lemmens, N. Van Roy, A. Soete, J. Reumers, M. Moisse, S. Plaisance, R. Drmanac, J. Chen, F. Speleman, D. Lambrechts, Y. Van de Peer, J. Tavernier, and N. Callewaert. Genome dynamics of the

- human embryonic kidney 293 lineage in response to cell biology manipulations. *Nature Communications*, 5(1):4767, dec 2014. ISSN 2041-1723. doi: 10.1038/ncomms5767. URL <http://www.nature.com/articles/ncomms5767>.
- [146] S. M. Linker, L. Urban, S. J. Clark, M. Chhatriwala, S. Amatya, D. J. McCarthy, I. Ebersberger, L. Vallier, W. Reik, O. Stegle, and M. J. Bonder. Combined single-cell profiling of expression and DNA methylation reveals splicing regulation and heterogeneity. *Genome Biology*, 20(1):30, dec 2019. ISSN 1474-760X. doi: 10.1186/s13059-019-1644-0. URL <https://genomebiology.biomedcentral.com/articles/10.1186/s13059-019-1644-0>.
- [147] R. Lister, M. Pelizzola, R. H. Downen, R. D. Hawkins, G. Hon, J. Tonti-Filippini, J. R. Nery, L. Lee, Z. Ye, Q.-M. Ngo, L. Edsall, J. Antosiewicz-Bourget, R. Stewart, V. Ruotti, A. H. Millar, J. A. Thomson, B. Ren, and J. R. Ecker. Human DNA methylomes at base resolution show widespread epigenomic differences. *Nature*, 462(7271):315–322, nov 2009. ISSN 0028-0836. doi: 10.1038/nature08514. URL <http://www.nature.com/articles/nature08514>.
- [148] W.-M. Liu and C. W. Schmid. Proposed roles for DNA methylation in Alu transcriptional repression and mutational inactivation. *Nucleic Acids Research*, 21(6):1351–1359, 1993. ISSN 0305-1048. doi: 10.1093/nar/21.6.1351. URL <https://academic.oup.com/nar/article-lookup/doi/10.1093/nar/21.6.1351>.
- [149] X. Liu, J. Freitas, D. Zheng, M. S. Oliveira, M. Hoque, T. Martins, T. Henriques, B. Tian, and A. Moreira. Transcription elongation rate has a tissue-specific impact on alternative cleavage and polyadenylation in *Drosophila melanogaster*. *RNA*, 23(12):1807–1816, dec 2017. ISSN 1355-8382. doi: 10.1261/rna.062661.117. URL <http://rnajournal.cshlp.org/lookup/doi/10.1261/rna.062661.117>.
- [150] Y. Liu, P. Siejka-Zielińska, G. Velikova, Y. Bi, F. Yuan, M. Tomkova, C. Bai, L. Chen, B. Schuster-Böckler, and C.-X. Song. Bisulfite-free direct detection of 5-methylcytosine and 5-hydroxymethylcytosine at base resolution. *Nature Biotechnology*, 37(4):424–429, apr 2019. ISSN 1087-0156.

doi: 10.1038/s41587-019-0041-2. URL <http://www.nature.com/articles/s41587-019-0041-2>.

- [151] Y. Liu, Z. Hu, J. Cheng, P. Siejka-Zielińska, J. Chen, M. Inoue, A. A. Ahmed, and C.-X. Song. Subtraction-free and bisulfite-free specific sequencing of 5-methylcytosine and its oxidized derivatives at base resolution. *Nature Communications*, 12(1):618, dec 2021. ISSN 2041-1723. doi: 10.1038/s41467-021-20920-2. URL <http://www.nature.com/articles/s41467-021-20920-2>.
- [152] M. C. Lorincz, D. R. Dickerson, M. Schmitt, and M. Groudine. Intragenic DNA methylation alters chromatin structure and elongation efficiency in mammalian cells. *Nature Structural & Molecular Biology*, 11(11):1068–1075, nov 2004. ISSN 1545-9993. doi: 10.1038/nsmb840. URL <http://www.nature.com/articles/nsmb840>.
- [153] M. I. Love, W. Huber, and S. Anders. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biology*, 15(12):550, dec 2014. ISSN 1474-760X. doi: 10.1186/s13059-014-0550-8. URL <http://genomebiology.biomedcentral.com/articles/10.1186/s13059-014-0550-8>.
- [154] M. I. Love, J. B. Hogenesch, and R. A. Irizarry. Modeling of RNA-seq fragment sequence bias reduces systematic errors in transcript abundance estimation. *Nature Biotechnology*, 34(12):1287–1291, dec 2016. ISSN 1087-0156. doi: 10.1038/nbt.3682. URL <http://www.nature.com/articles/nbt.3682>.
- [155] C. Lövkvist, I. B. Dodd, K. Sneppen, and J. O. Haerter. DNA methylation in human epigenomes depends on local topology of CpG sites. *Nucleic Acids Research*, 44(11):5123–5132, jun 2016. ISSN 0305-1048. doi: 10.1093/nar/gkw124. URL <https://academic.oup.com/nar/article-lookup/doi/10.1093/nar/gkw124>.
- [156] D. S. Luse, M. Parida, B. M. Spector, K. A. Nilson, and D. H. Price. A unified view of the sequence and functional organization of the human RNA polymerase II promoter. *Nucleic Acids Research*, 48(14):7767–7785, aug 2020. ISSN 0305-1048.

doi: 10.1093/nar/gkaa531. URL <https://academic.oup.com/nar/article/48/14/7767/5864710>.

- [157] J. Madzo, H. Liu, A. Rodriguez, A. Vasanthakumar, S. Sundaravel, D. B. D. Caces, T. J. Looney, L. Zhang, J. B. Lepore, T. Macrae, R. Duszynski, A. H. Shih, C.-X. Song, M. Yu, Y. Yu, R. Grossman, B. Raumann, A. Verma, C. He, R. L. Levine, D. Lavelle, B. T. Lahn, A. Wickrema, and L. A. Godley. Hydroxymethylation at Gene Regulatory Regions Directs Stem/Early Progenitor Cell Commitment during Erythropoiesis. *Cell Reports*, 6(1):231–244, jan 2014. ISSN 22111247. doi: 10.1016/j.celrep.2013.11.044. URL <https://linkinghub.elsevier.com/retrieve/pii/S2211124713007286>.
- [158] D. B. Mahat and J. T. Lis. Use of conditioned media is critical for studies of regulation in response to rapid heat shock. *Cell Stress and Chaperones*, 22(1):155–162, jan 2017. ISSN 1355-8145. doi: 10.1007/s12192-016-0737-x. URL <http://link.springer.com/10.1007/s12192-016-0737-x>.
- [159] D. B. Mahat, H. Kwak, G. T. Booth, I. H. Jonkers, C. G. Danko, R. K. Patel, C. T. Waters, K. Munson, L. J. Core, and J. T. Lis. Base-pair-resolution genome-wide mapping of active RNA polymerases using precision nuclear run-on (PRO-seq). *Nature Protocols*, 11(8):1455–1476, aug 2016. ISSN 1754-2189. doi: 10.1038/nprot.2016.086. URL <https://www.ncbi.nlm.nih.gov/pubmed/27442863http://www.nature.com/articles/nprot.2016.086>.
- [160] B. Maher. Personal genomes: The case of the missing heritability. *Nature*, 456(7218):18–21, nov 2008. ISSN 0028-0836. doi: 10.1038/456018a. URL <http://www.nature.com/articles/456018a>.
- [161] A. Maiti and A. C. Drohat. Thymine DNA Glycosylase Can Rapidly Excise 5-Formylcytosine and 5-Carboxylcytosine. *Journal of Biological Chemistry*, 286(41):35334–35338, oct 2011. ISSN 00219258. doi: 10.1074/jbc.C111.284620. URL <https://linkinghub.elsevier.com/retrieve/pii/S0021925820737385>.
- [162] P. B. Mason and K. Struhl. Distinction and Relationship between Elongation

- Rate and Processivity of RNA Polymerase II In Vivo. *Molecular Cell*, 17(6): 831–840, mar 2005. ISSN 10972765. doi: 10.1016/j.molcel.2005.02.017. URL <https://linkinghub.elsevier.com/retrieve/pii/S1097276505011160>.
- [163] A. K. Maunakea, I. Chepelev, K. Cui, and K. Zhao. Intragenic DNA methylation modulates alternative splicing by recruiting MeCP2 to promote exon recognition. *Cell Research*, 23(11):1256–1269, nov 2013. ISSN 1001-0602. doi: 10.1038/cr.2013.110. URL <http://www.nature.com/articles/cr2013110>.
- [164] J. McDowell, J. Roberts, D. Jin, and C. Gross. Determination of intrinsic transcription termination efficiency by RNA polymerase elongation rate. *Science*, 266(5186):822–825, nov 1994. ISSN 0036-8075. doi: 10.1126/science.7526463. URL <https://www.sciencemag.org/lookup/doi/10.1126/science.7526463>.
- [165] M. Mellén, P. Ayata, S. Dewell, S. Kriaucionis, and N. Heintz. MeCP2 Binds to 5hmC Enriched within Active Genes and Accessible Chromatin in the Nervous System. *Cell*, 151(7):1417–1430, dec 2012. ISSN 00928674. doi: 10.1016/j.cell.2012.11.022. URL <https://linkinghub.elsevier.com/retrieve/pii/S0092867412014079>.
- [166] B. Miller, F. Sánchez-Vega, and L. Elnitski. The Emergence of Pan-Cancer CIMP and Its Elusive Interpretation. *Biomolecules*, 6(4):45, nov 2016. ISSN 2218-273X. doi: 10.3390/biom6040045. URL <http://www.mdpi.com/2218-273X/6/4/45>.
- [167] K. N. Mohni, S. R. Wessel, R. Zhao, A. C. Wojciechowski, J. W. Luzwick, H. Layden, B. F. Eichman, P. S. Thompson, K. P. Mehta, and D. Cortez. HMCES Maintains Genome Integrity by Shielding Abasic Sites in Single-Strand DNA. *Cell*, 176(1-2):144–153.e13, jan 2019. ISSN 00928674. doi: 10.1016/j.cell.2018.10.055. URL <https://linkinghub.elsevier.com/retrieve/pii/S0092867418314545>.
- [168] M. J. Moore and N. J. Proudfoot. Pre-mRNA Processing Reaches Back to Transcription and Ahead to Translation. *Cell*, 136(4):688–700, feb 2009. ISSN 00928674. doi: 10.1016/j.cell.2009.02.001. URL <https://linkinghub.elsevier.com/retrieve/pii/S0092867409001330>.

- [169] N. J. Mullen and D. H. Price. Hydrogen peroxide yields mechanistic insights into human mRNA capping enzyme function. *PLOS ONE*, 12(10):e0186423, oct 2017. ISSN 1932-6203. doi: 10.1371/journal.pone.0186423. URL <https://dx.plos.org/10.1371/journal.pone.0186423>.
- [170] S. C. Murray, S. Haenni, F. S. Howe, H. Fischl, K. Chocian, A. Nair, and J. Mellor. Sense and antisense transcription are associated with distinct chromatin architectures across genes. *Nucleic Acids Research*, 43(16):7823–7837, sep 2015. ISSN 0305-1048. doi: 10.1093/nar/gkv666. URL <https://academic.oup.com/nar/article-lookup/doi/10.1093/nar/gkv666>.
- [171] A. Nagari, S. Murakami, V. S. Malladi, and W. L. Kraus. Computational Approaches for Mining GRO-Seq Data to Identify and Characterize Active Enhancers. In *Methods in Molecular Biology*, volume 1468, pages 121–138. 2017. ISBN 978-1-4939-4033-2. doi: 10.1007/978-1-4939-4035-6_10. URL http://link.springer.com/10.1007/978-1-4939-4035-6_{_}10.
- [172] F. Neri, S. Rapelli, A. Krepelova, D. Incarnato, C. Parlato, G. Basile, M. Maldotti, F. Anselmi, and S. Oliviero. Intragenic DNA methylation prevents spurious transcription initiation. *Nature*, 543(7643):72–77, mar 2017. ISSN 0028-0836. doi: 10.1038/nature21373. URL <http://www.nature.com/articles/nature21373>.
- [173] J. Newell-Price, A. J. Clark, and P. King. DNA Methylation and Silencing of Gene Expression. *Trends in Endocrinology & Metabolism*, 11(4):142–148, may 2000. ISSN 10432760. doi: 10.1016/S1043-2760(00)00248-4. URL <https://linkinghub.elsevier.com/retrieve/pii/S1043276000002484>.
- [174] T. Nguyen, H. Fischl, F. S. Howe, R. Woloszczuk, A. Serra Barros, Z. Xu, D. Brown, S. C. Murray, S. Haenni, J. M. Halstead, L. O’Connor, G. Shipkovenska, L. M. Steinmetz, and J. Mellor. Transcription mediated insulation and interference direct gene cluster expression switches. *eLife*, 3, nov 2014. ISSN 2050-084X. doi: 10.7554/eLife.03635. URL <https://elifesciences.org/articles/03635>.
- [175] S. Noguchi, T. Arakawa, S. Fukuda, M. Furuno, A. Hasegawa, F. Hori,

S. Ishikawa-Kato, K. Kaida, A. Kaiho, M. Kanamori-Katayama, T. Kawashima, M. Kojima, A. Kubosaki, R.-i. Manabe, M. Murata, S. Nagao-Sato, K. Nakazato, N. Ninomiya, H. Nishiyori-Sueki, S. Noma, E. Saijyo, A. Saka, M. Sakai, C. Simon, N. Suzuki, M. Tagami, S. Watanabe, S. Yoshida, P. Arner, R. A. Axton, M. Babina, J. K. Baillie, T. C. Barnett, A. G. Beckhouse, A. Blumenthal, B. Bodega, A. Bonetti, J. Briggs, F. Brombacher, A. J. Carlisle, H. C. Clevers, C. A. Davis, M. Detmar, T. Dohi, A. S. Edge, M. Edinger, A. Ehrlund, K. Ekwall, M. Endoh, H. Enomoto, A. Eslami, M. Fagiolini, L. Fairbairn, M. C. Farach-Carson, G. J. Faulkner, C. Ferrai, M. E. Fisher, L. M. Forrester, R. Fujita, J.-i. Furusawa, T. B. Geijtenbeek, T. Gingeras, D. Goldowitz, S. Guhl, R. Guler, S. Gustincich, T. J. Ha, M. Hamaguchi, M. Hara, Y. Hasegawa, M. Herlyn, P. Heutink, K. J. Hitchens, D. A. Hume, T. Ikawa, Y. Ishizu, C. Kai, H. Kawamoto, Y. I. Kawamura, J. S. Kempfle, T. J. Kenna, J. Kere, L. M. Khachigian, T. Kitamura, S. Klein, S. P. Klinken, A. J. Knox, S. Kojima, H. Koseki, S. Koyasu, W. Lee, A. Lennartsson, A. Mackay-sim, N. Mejhert, Y. Mizuno, H. Morikawa, M. Morimoto, K. Moro, K. J. Morris, H. Motohashi, C. L. Mummery, Y. Nakachi, F. Nakahara, T. Nakamura, Y. Nakamura, T. Nozaki, S. Ogishima, N. Ohkura, H. Ohno, M. Ohshima, M. Okada-Hatakeyama, Y. Okazaki, V. Orlando, D. A. Ovchinnikov, R. Passier, M. Patrikakis, A. Pombo, S. Pradhan-Bhatt, X.-Y. Qin, M. Rehli, P. Rizzu, S. Roy, A. Sajantila, S. Sakaguchi, H. Sato, H. Satoh, S. Savvi, A. Saxena, C. Schmidl, C. Schneider, G. G. Schulze-Tanzil, A. Schwegmann, G. Sheng, J. W. Shin, D. Sugiyama, T. Sugiyama, K. M. Summers, N. Takahashi, J. Takai, H. Tanaka, H. Tatsukawa, A. Tomoiu, H. Toyoda, M. van de Wetering, L. M. van den Berg, R. Verardo, D. Vijayan, C. A. Wells, L. N. Winteringham, E. Wolvetang, Y. Yamaguchi, M. Yamamoto, C. Yanagi-Mizuochi, M. Yoneda, Y. Yonekura, P. G. Zhang, S. Zucchelli, I. Abugessaisa, E. Arner, J. Harshbarger, A. Kondo, T. Lassmann, M. Lizio, S. Sahin, T. Sengstag, J. Severin, H. Shimoji, M. Suzuki, H. Suzuki, J. Kawai, N. Kondo, M. Itoh, C. O. Daub, T. Kasukawa, H. Kawaji, P. Carninci, A. R. Forrest, and Y. Hayashizaki. FANTOM5 CAGE profiles of

- human and mouse samples. *Scientific Data*, 4(1):170112, dec 2017. ISSN 2052-4463. doi: 10.1038/sdata.2017.112. URL <http://www.nature.com/articles/sdata2017112>.
- [176] T. Nojima, T. Gomes, A. R. F. Grosso, H. Kimura, M. J. Dye, S. Dhir, M. Carmo-Fonseca, and N. J. Proudfoot. Mammalian NET-Seq Reveals Genome-wide Nascent Transcription Coupled to RNA Processing. *Cell*, 161(3):526–540, apr 2015. ISSN 00928674. doi: 10.1016/j.cell.2015.03.027. URL <https://linkinghub.elsevier.com/retrieve/pii/S0092867415003141>.
- [177] M. Okano, D. W. Bell, D. A. Haber, and E. Li. DNA Methyltransferases Dnmt3a and Dnmt3b Are Essential for De Novo Methylation and Mammalian Development. *Cell*, 99(3):247–257, oct 1999. ISSN 00928674. doi: 10.1016/S0092-8674(00)81656-6. URL <https://linkinghub.elsevier.com/retrieve/pii/S0092867400816566>.
- [178] I. Papathanasiou, V. Trachana, E. Mourmoura, and A. Tsezou. DNA methylation regulates miR-140-5p and miR-146a expression in osteoarthritis. *Life Sciences*, 228:274–284, jul 2019. ISSN 00243205. doi: 10.1016/j.lfs.2019.05.018. URL <https://linkinghub.elsevier.com/retrieve/pii/S0024320519303649>.
- [179] M. Parida, K. A. Nilson, M. Li, C. B. Ball, H. A. Fuchs, C. K. Lawson, D. S. Luse, J. L. Meier, and D. H. Price. Nucleotide Resolution Comparison of Transcription of Human Cytomegalovirus and Host Genomes Reveals Universal Use of RNA Polymerase II Elongation Control Driven by Dissimilar Core Promoter Elements. *mBio*, 10(1), feb 2019. ISSN 2161-2129. doi: 10.1128/mBio.02047-18. URL <https://journals.asm.org/doi/10.1128/mBio.02047-18>.
- [180] P. K. Parua, G. T. Booth, M. Sansó, B. Benjamin, J. C. Tanny, J. T. Lis, and R. P. Fisher. A Cdk9–PP1 switch regulates the elongation–termination transition of RNA polymerase II. *Nature*, 558(7710):460–464, jun 2018. ISSN 0028-0836. doi: 10.1038/s41586-018-0214-z. URL <http://www.nature.com/articles/s41586-018-0214-z>.

- [181] W. A. Pastor, U. J. Pape, Y. Huang, H. R. Henderson, R. Lister, M. Ko, E. M. McLoughlin, Y. Brudno, S. Mahapatra, P. Kapranov, M. Tahiliani, G. Q. Daley, X. S. Liu, J. R. Ecker, P. M. Milos, S. Agarwal, and A. Rao. Genome-wide mapping of 5-hydroxymethylcytosine in embryonic stem cells. *Nature*, 473(7347): 394–397, may 2011. ISSN 0028-0836. doi: 10.1038/nature10102. URL <http://www.nature.com/articles/nature10102>.
- [182] V. Patil, R. L. Ward, and L. B. Hesson. The evidence for functional non-CpG methylation in mammalian cells. *Epigenetics*, 9(6):823–828, jun 2014. ISSN 1559-2294. doi: 10.4161/epi.28741. URL <http://www.tandfonline.com/doi/abs/10.4161/epi.28741>.
- [183] V. Patil, C. Cuenin, F. Chung, J. R. Aguilera, N. Fernandez-Jimenez, I. Romero-Garmendia, J. R. Bilbao, V. Cahais, J. Rothwell, and Z. Herceg. Human mitochondrial DNA is extensively methylated in a non-CpG context. *Nucleic acids research*, 2019. ISSN 13624962. doi: 10.1093/nar/gkz762.
- [184] R. Patro, G. Duggal, M. I. Love, R. A. Irizarry, and C. Kingsford. Salmon provides fast and bias-aware quantification of transcript expression. *Nature Methods*, 14(4):417–419, apr 2017. ISSN 1548-7091. doi: 10.1038/nmeth.4197. URL <http://www.nature.com/articles/nmeth.4197>.
- [185] V. Pelechano and L. M. Steinmetz. (REVIEW) Gene regulation by antisense transcription, 2013. ISSN 14710056.
- [186] T. Pfaffeneder, B. Hackner, M. Truß, M. Münzel, M. Müller, C. A. Deiml, C. Hagemeier, and T. Carell. The Discovery of 5-Formylcytosine in Embryonic Stem Cell DNA. *Angewandte Chemie International Edition*, 50(31):7008–7012, jul 2011. ISSN 14337851. doi: 10.1002/anie.201103899. URL <http://doi.wiley.com/10.1002/anie.201103899>.
- [187] A. Pirola-Merlo, C. Härtel, L. Mann, and G. Hirst. How leaders influence the impact of affective events on team climate and performance in R&D teams. *The Leadership Quarterly*, 13(5):561–581, oct 2002. ISSN 10489843. doi: 10.1016/

- S1048-9843(02)00144-3. URL <https://linkinghub.elsevier.com/retrieve/pii/S1048984302001443>.
- [188] N. J. Proudfoot. (REVIEW) Transcriptional termination in mammals: Stopping the RNA polymerase II juggernaut, 2016. ISSN 10959203.
- [189] E. L. Putiri and K. D. Robertson. Epigenetic mechanisms and genome stability. *Clinical Epigenetics*, 2(2):299–314, aug 2011. ISSN 1868-7075. doi: 10.1007/s13148-010-0017-z. URL <http://www.clinicalepigeneticsjournal.com/content/2/2/>.
- [190] A. R. Quinlan and I. M. Hall. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics*, 26(6):841–842, mar 2010. ISSN 1460-2059. doi: 10.1093/bioinformatics/btq033. URL <https://academic.oup.com/bioinformatics/article-lookup/doi/10.1093/bioinformatics/btq033>.
- [191] R Core Team. R: A language and environment for statistical computing., 2019.
- [192] P. B. Rahl, C. Y. Lin, A. C. Seila, R. A. Flynn, S. McCuine, C. B. Burge, P. A. Sharp, and R. A. Young. c-Myc Regulates Transcriptional Pause Release. *Cell*, 141(3):432–445, apr 2010. ISSN 00928674. doi: 10.1016/j.cell.2010.03.030. URL <https://linkinghub.elsevier.com/retrieve/pii/S0092867410003181>.
- [193] E.-A. Raiber, P. Murat, D. Y. Chirgadze, D. Beraldi, B. F. Luisi, and S. Balasubramanian. 5-Formylcytosine alters the structure of the DNA double helix. *Nature Structural & Molecular Biology*, 22(1):44–49, jan 2015. ISSN 1545-9993. doi: 10.1038/nsmb.2936. URL <http://www.nature.com/articles/nsmb.2936>.
- [194] A. Ramassone, S. Pagotto, A. Veronese, and R. Visone. Epigenetics and MicroRNAs in Cancer. *International Journal of Molecular Sciences*, 19(2):459, feb 2018. ISSN 1422-0067. doi: 10.3390/ijms19020459. URL <http://www.mdpi.com/1422-0067/19/2/459>.
- [195] A. C. Rand, M. Jain, J. M. Eizenga, A. Musselman-Brown, H. E. Olsen, M. Akeson, and B. Paten. Mapping DNA methylation with high-throughput nanopore

- sequencing. *Nature Methods*, 14(4):411–413, apr 2017. ISSN 1548-7091. doi: 10.1038/nmeth.4189. URL <http://www.nature.com/articles/nmeth.4189>.
- [196] J. Ray, P. R. Munn, A. Vihervaara, J. J. Lewis, A. Ozer, C. G. Danko, and J. T. Lis. Chromatin conformation remains stable upon extensive transcriptional changes driven by heat shock. *Proceedings of the National Academy of Sciences*, 116(39):19431–19439, sep 2019. ISSN 0027-8424. doi: 10.1073/pnas.1901244116. URL <http://www.pnas.org/lookup/doi/10.1073/pnas.1901244116>.
- [197] I. Rhee, K. E. Bachman, B. H. Park, K.-W. Jair, R.-W. C. Yen, K. E. Schuebel, H. Cui, A. P. Feinberg, C. Lengauer, K. W. Kinzler, S. B. Baylin, and B. Vogelstein. DNMT1 and DNMT3b cooperate to silence genes in human cancer cells. *Nature*, 416(6880):552–556, apr 2002. ISSN 0028-0836. doi: 10.1038/416552a. URL <http://www.nature.com/articles/416552a>.
- [198] R. Rickels, H.-M. Herz, C. C. Sze, K. Cao, M. A. Morgan, C. K. Collings, M. Gause, Y.-h. Takahashi, L. Wang, E. J. Rendleman, S. A. Marshall, A. Krueger, E. T. Bartom, A. Piunti, E. R. Smith, N. A. Abshiru, N. L. Kelleher, D. Dorsett, and A. Shilatifard. Histone H3K4 monomethylation catalyzed by Trr and mammalian COMPASS-like proteins at enhancers is dispensable for development and viability. *Nature Genetics*, 49(11):1647–1653, nov 2017. ISSN 1061-4036. doi: 10.1038/ng.3965. URL <http://www.nature.com/articles/ng.3965>.
- [199] A. Riggs. X inactivation, differentiation, and DNA methylation. *Cytogenetic and Genome Research*, 14(1):9–25, 1975. ISSN 1424-859X. doi: 10.1159/000130315. URL <https://www.karger.com/Article/FullText/130315>.
- [200] M.-F. Robert, S. Morin, N. Beaulieu, F. Gauthier, I. C. Chute, A. Barsalou, and A. R. MacLeod. DNMT1 is required to maintain CpG methylation and aberrant gene silencing in human cancer cells. *Nature Genetics*, 33(1):61–65, jan 2003. ISSN 1061-4036. doi: 10.1038/ng1068. URL <http://www.nature.com/articles/ng1068z>.
- [201] A. Roberts, C. Trapnell, J. Donaghey, J. L. Rinn, and L. Pachter. Improving

- RNA-Seq expression estimates by correcting for fragment bias. *Genome Biology*, 12(3):R22, 2011. ISSN 1465-6906. doi: 10.1186/gb-2011-12-3-r22. URL <http://genomebiology.biomedcentral.com/articles/10.1186/gb-2011-12-3-r22>.
- [202] A. B. Robertson, J. Robertson, M. Fusser, and A. Klungland. Endonuclease G preferentially cleaves 5-hydroxymethylcytosine-modified DNA creating a substrate for recombination. *Nucleic Acids Research*, 42(21):13280–13293, dec 2014. ISSN 1362-4962. doi: 10.1093/nar/gku1032. URL <http://academic.oup.com/nar/article/42/21/13280/2902637/Endonuclease-G-preferentially-cleaves>.
- [203] J. Robertson, A. B. Robertson, and A. Klungland. The presence of 5-hydroxymethylcytosine at the gene promoter and not in the gene body negatively regulates gene expression. *Biochemical and Biophysical Research Communications*, 411(1):40–43, jul 2011. ISSN 0006291X. doi: 10.1016/j.bbrc.2011.06.077. URL <https://linkinghub.elsevier.com/retrieve/pii/S0006291X11010503>.
- [204] M. Santiago, C. Antunes, M. Guedes, N. Sousa, and C. J. Marques. TET enzymes and DNA hydroxymethylation in neural development and function — How critical are they? *Genomics*, 104(5):334–340, nov 2014. ISSN 08887543. doi: 10.1016/j.ygeno.2014.08.018. URL <https://linkinghub.elsevier.com/retrieve/pii/S0888754314001608>.
- [205] L. A. Sanz, S. R. Hartono, Y. W. Lim, S. Steyaert, A. Rajpurkar, P. A. Ginno, X. Xu, and F. Chédin. Prevalent, Dynamic, and Conserved R-Loop Structures Associate with Specific Epigenomic Signatures in Mammals. *Molecular Cell*, 63(1):167–178, jul 2016. ISSN 10972765. doi: 10.1016/j.molcel.2016.05.032. URL <https://linkinghub.elsevier.com/retrieve/pii/S1097276516301964>.
- [206] M. Sasaki, C. B. Knobbe, J. C. Munger, E. F. Lind, D. Brenner, A. Brüstle, I. S. Harris, R. Holmes, A. Wakeham, J. Haight, A. You-Ten, W. Y. Li, S. Schalm, S. M. Su, C. Virtanen, G. Reifemberger, P. S. Ohashi, D. L. Barber, M. E. Figueroa, A. Melnick, J.-C. Zúñiga-Pflücker, and T. W. Mak. IDH1(R132H) mutation increases murine haematopoietic progenitors and alters epigenetics. *Nature*, 488

- (7413):656–659, aug 2012. ISSN 0028-0836. doi: 10.1038/nature11323. URL <http://www.nature.com/articles/nature11323>.
- [207] S. Saxonov, P. Berg, and D. L. Brutlag. A genome-wide analysis of CpG dinucleotides in the human genome distinguishes two distinct classes of promoters. *Proceedings of the National Academy of Sciences*, 103(5):1412–1417, jan 2006. ISSN 0027-8424. doi: 10.1073/pnas.0510310103. URL <http://www.pnas.org/cgi/doi/10.1073/pnas.0510310103>.
- [208] K. Schaukowitch, J.-Y. Joo, X. Liu, J. K. Watts, C. Martinez, and T.-K. Kim. Enhancer RNA Facilitates NELF Release from Immediate Early Genes. *Molecular Cell*, 56(1):29–42, oct 2014. ISSN 10972765. doi: 10.1016/j.molcel.2014.08.023. URL <https://linkinghub.elsevier.com/retrieve/pii/S1097276514006789>.
- [209] F. Schlesinger, A. D. Smith, T. R. Gingeras, G. J. Hannon, and E. Hodges. De novo DNA demethylation and noncoding transcription define active intergenic regulatory elements. *Genome Research*, 2013. ISSN 10889051. doi: 10.1101/gr.157271.113.
- [210] J. Schreiber. pomegranate: Fast and flexible probabilistic modeling in python. *Journal of Machine Learning Research*, 18(164):1–6, 2018. ISSN 15337928.
- [211] D. Schübeler. Function and information content of DNA methylation. *Nature*, 517(7534):321–326, jan 2015. ISSN 0028-0836. doi: 10.1038/nature14192. URL <http://www.nature.com/articles/nature14192>.
- [212] B. Schwalb, M. Michel, B. Zacher, K. Frühauf, C. Demel, A. Tresch, J. Gagneur, and P. Cramer. TT-seq maps the human transient transcriptome. *Science*, 352(6290):1225–1228, jun 2016. ISSN 0036-8075. doi: 10.1126/science.aad9841. URL <https://www.sciencemag.org/lookup/doi/10.1126/science.aad9841>.
- [213] B. S. Scruggs, D. A. Gilchrist, S. Nechaev, G. W. Muse, A. Burkholder, D. C. Fargo, and K. Adelman. Bidirectional Transcription Arises from Two Distinct Hubs of Transcription Factor Binding and Active Chromatin. *Molecular Cell*, 58(6):1101–1112, 2015. ISSN 10974164. doi: 10.1016/j.molcel.2015.04.006.

- [214] C. L. Seiler, J. Fernandez, Z. Koerperich, M. P. Andersen, D. Kotandeniya, M. E. Nguyen, Y. Y. Sham, and N. Y. Tretyakova. Maintenance DNA Methyltransferase Activity in the Presence of Oxidized Forms of 5-Methylcytosine: Structural Basis for Ten Eleven Translocation-Mediated DNA Demethylation. *Biochemistry*, 57(42):6061–6069, oct 2018. ISSN 0006-2960. doi: 10.1021/acs.biochem.8b00683. URL <https://pubs.acs.org/doi/10.1021/acs.biochem.8b00683>.
- [215] R. Shapiro, R. E. Servis, and M. Welcher. Reactions of Uracil and Cytosine Derivatives with Sodium Bisulfite. *Journal of the American Chemical Society*, 92(2):422–424, jan 1970. ISSN 0002-7863. doi: 10.1021/ja00705a626. URL <https://pubs.acs.org/doi/abs/10.1021/ja00705a626>.
- [216] A. Sharifi-Zarchi, D. Gerovska, K. Adachi, M. Totonchi, H. Pezeshk, R. J. Taft, H. R. Schöler, H. Chitsaz, M. Sadeghi, H. Baharvand, and M. J. Araújo-Bravo. DNA methylation regulates discrimination of enhancers from promoters through a H3K4me1-H3K4me3 seesaw mechanism. *BMC Genomics*, 18(1):964, dec 2017. ISSN 1471-2164. doi: 10.1186/s12864-017-4353-7. URL <https://bmcbgenomics.biomedcentral.com/articles/10.1186/s12864-017-4353-7>.
- [217] A. J. Sharp, E. Stathaki, E. Migliavacca, M. Brahmachary, S. B. Montgomery, Y. Dupre, and S. E. Antonarakis. DNA methylation profiles of human active and inactive X chromosomes. *Genome Research*, 21(10):1592–1600, oct 2011. ISSN 1088-9051. doi: 10.1101/gr.112680.110. URL <http://genome.cshlp.org/cgi/doi/10.1101/gr.112680.110>.
- [218] R. Shayevitch, D. Askayo, I. Keydar, and G. Ast. The importance of DNA methylation of exons on alternative splicing. *RNA*, 24(10):1351–1362, oct 2018. ISSN 1355-8382. doi: 10.1261/rna.064865.117. URL <http://rnajournal.cshlp.org/lookup/doi/10.1261/rna.064865.117>.
- [219] V. Shukla, L. Halabelian, S. Balagere, D. Samaniego-Castruita, D. E. Feldman, C. H. Arrowsmith, A. Rao, and L. Aravind. HMCES Functions in the Alternative End-Joining Pathway of the DNA DSB Repair during Class Switch Recombination in B Cells. *Molecular Cell*, 77(2):384–394.e4, jan 2020. ISSN 10972765.

- doi: 10.1016/j.molcel.2019.10.031. URL <https://linkinghub.elsevier.com/retrieve/pii/S1097276519308093>.
- [220] T. Smith, A. Heger, and I. Sudbery. UMI-tools: modeling sequencing errors in Unique Molecular Identifiers to improve quantification accuracy. *Genome Research*, 27(3):491–499, mar 2017. ISSN 1088-9051. doi: 10.1101/gr.209601.116. URL <http://genome.cshlp.org/lookup/doi/10.1101/gr.209601.116>.
- [221] C. Soneson, M. I. Love, and M. D. Robinson. Differential analyses for RNA-seq: transcript-level estimates improve gene-level inferences. *F1000Research*, 4:1521, dec 2015. ISSN 2046-1402. doi: 10.12688/f1000research.7563.1. URL <https://f1000research.com/articles/4-1521/v1>.
- [222] C.-X. Song, K. E. Szulwach, Y. Fu, Q. Dai, C. Yi, X. Li, Y. Li, C.-H. Chen, W. Zhang, X. Jian, J. Wang, L. Zhang, T. J. Looney, B. Zhang, L. A. Godley, L. M. Hicks, B. T. Lahn, P. Jin, and C. He. Selective chemical labeling reveals the genome-wide distribution of 5-hydroxymethylcytosine. *Nature Biotechnology*, 29(1):68–72, jan 2011. ISSN 1087-0156. doi: 10.1038/nbt.1732. URL <http://www.nature.com/articles/nbt.1732>.
- [223] C.-X. Song, K. E. Szulwach, Q. Dai, Y. Fu, S.-Q. Mao, L. Lin, C. Street, Y. Li, M. Poidevin, H. Wu, J. Gao, P. Liu, L. Li, G.-L. Xu, P. Jin, and C. He. Genome-wide Profiling of 5-Formylcytosine Reveals Its Roles in Epigenetic Priming. *Cell*, 153(3):678–691, apr 2013. ISSN 00928674. doi: 10.1016/j.cell.2013.04.001. URL <https://linkinghub.elsevier.com/retrieve/pii/S0092867413004005>.
- [224] C.-X. Song, J. Diao, A. T. Brunger, and S. R. Quake. Simultaneous single-molecule epigenetic imaging of DNA methylation and hydroxymethylation. *Proceedings of the National Academy of Sciences*, 113(16):4338–4343, apr 2016. ISSN 0027-8424. doi: 10.1073/pnas.1600223113. URL <http://www.pnas.org/lookup/doi/10.1073/pnas.1600223113>.
- [225] M. B. Stadler, R. Murr, L. Burger, R. Ivanek, F. Lienert, A. Schöler, E. van Nimwegen, C. Wirbelauer, E. J. Oakeley, D. Gaidatzis, V. K. Tiwari, and

- D. Schübeler. DNA-binding factors shape the mouse methylome at distal regulatory regions. *Nature*, 480(7378):490–495, dec 2011. ISSN 0028-0836. doi: 10.1038/nature10716. URL <http://www.nature.com/articles/nature10716>.
- [226] I. Steinparzer, V. Sedlyarov, J. D. Rubin, K. Eislmayr, M. D. Galbraith, C. B. Levandowski, T. Vcelkova, L. Sneezum, F. Wascher, F. Amman, R. Kleinova, H. Bender, Z. Andrysik, J. M. Espinosa, G. Superti-Furga, R. D. Dowell, D. J. Taatjes, and P. Kovarik. Transcriptional Responses to IFN- γ Require Mediator Kinase-Dependent Pause Release and Mechanistically Distinct CDK8 and CDK19 Functions. *Molecular Cell*, 76(3):485–499.e8, nov 2019. ISSN 10972765. doi: 10.1016/j.molcel.2019.07.034. URL <https://linkinghub.elsevier.com/retrieve/pii/S1097276519305891>.
- [227] A. Stepanenko and V. Dmitrenko. HEK293 in cell biology and cancer research: phenotype, karyotype, tumorigenicity, and stress-induced genome-phenotype evolution. *Gene*, 569(2):182–190, sep 2015. ISSN 03781119. doi: 10.1016/j.gene.2015.05.065. URL <https://linkinghub.elsevier.com/retrieve/pii/S0378111915006502>.
- [228] B. Steurer, R. C. Janssens, B. Geverts, M. E. Geijer, F. Wienholz, A. F. Theil, J. Chang, S. Dealy, J. Pothof, W. A. van Cappellen, A. B. Houtsmuller, and J. A. Marteiijn. Live-cell analysis of endogenous GFP-RPB1 uncovers rapid turnover of initiating and promoter-paused RNA Polymerase II. *Proceedings of the National Academy of Sciences*, 2018. ISSN 0027-8424. doi: 10.1073/pnas.1717920115.
- [229] C. Stirzaker, J. Z. Song, W. Ng, Q. Du, N. J. Armstrong, W. J. Locke, A. L. Statham, H. French, R. Pidsley, F. Valdes-Mora, E. Zotenko, and S. J. Clark. Methyl-CpG-binding protein MBD2 plays a key role in maintenance and spread of DNA methylation at CpG islands and shores in cancer. *Oncogene*, 36(10):1328–1338, mar 2017. ISSN 0950-9232. doi: 10.1038/onc.2016.297. URL <http://www.nature.com/articles/onc2016297>.
- [230] D. Strbenac, N. J. Armstrong, and J. Y. Yang. Detection and classification of peaks in 5' cap RNA sequencing data. *BMC Genomics*, 14(Suppl 5):S9, 2013.

ISSN 1471-2164. doi: 10.1186/1471-2164-14-S5-S9. URL <http://bmcgenomics.biomedcentral.com/articles/10.1186/1471-2164-14-S5-S9>.

- [231] H. Stroud, S. Feng, S. Morey Kinney, S. Pradhan, and S. E. Jacobsen. 5-Hydroxymethylcytosine is associated with enhancers and gene bodies in human embryonic stem cells. *Genome Biology*, 12(6):R54, 2011. ISSN 1465-6906. doi: 10.1186/gb-2011-12-6-r54. URL <http://genomebiology.biomedcentral.com/articles/10.1186/gb-2011-12-6-r54>.
- [232] M. M. Suzuki and A. Bird. DNA methylation landscapes: provocative insights from epigenomics. *Nature Reviews Genetics*, 9(6):465–476, jun 2008. ISSN 1471-0056. doi: 10.1038/nrg2341. URL <http://www.nature.com/articles/nrg2341>.
- [233] K. E. Szulwach, X. Li, Y. Li, C. X. Song, H. Wu, Q. Dai, H. Irier, A. K. Upadhyay, M. Gearing, A. I. Levey, A. Vasanthakumar, L. A. Godley, Q. Chang, X. Cheng, C. He, and P. Jin. 5-hmC-mediated epigenetic dynamics during postnatal neurodevelopment and aging. *Nature Neuroscience*, 2011. ISSN 10976256. doi: 10.1038/nn.2959.
- [234] M. Tahiliani, K. P. Koh, Y. Shen, W. A. Pastor, H. Bandukwala, Y. Brudno, S. Agarwal, L. M. Iyer, D. R. Liu, L. Aravind, and A. Rao. Conversion of 5-Methylcytosine to 5-Hydroxymethylcytosine in Mammalian DNA by MLL Partner TET1. *Science*, 324(5929):930–935, may 2009. ISSN 0036-8075. doi: 10.1126/science.1170116. URL <https://www.sciencemag.org/lookup/doi/10.1126/science.1170116>.
- [235] N. TAKAGI and M. SASAKI. Preferential inactivation of the paternally derived X chromosome in the extraembryonic membranes of the mouse. *Nature*, 256(5519):640–642, aug 1975. ISSN 0028-0836. doi: 10.1038/256640a0. URL <http://www.nature.com/articles/256640a0>.
- [236] A. Teissandier and D. Bourc’his. Gene body <scp>DNA</scp> methylation conspires with H3K36me3 to preclude aberrant transcription. *The EMBO Journal*,

- 36(11):1471–1473, jun 2017. ISSN 0261-4189. doi: 10.15252/emj.201796812. URL <https://onlinelibrary.wiley.com/doi/10.15252/emj.201796812>.
- [237] M. Thodberg, A. Thieffry, K. Vitting-Seerup, R. Andersson, and A. Sandelin. CAGEfightR: analysis of 5-end data using R/Bioconductor. *BMC Bioinformatics*, 20(1):487, dec 2019. ISSN 1471-2105. doi: 10.1186/s12859-019-3029-5. URL <https://bmcbioinformatics.biomedcentral.com/articles/10.1186/s12859-019-3029-5>.
- [238] J. M. Tome, N. D. Tippens, and J. T. Lis. Single-molecule nascent RNA sequencing identifies regulatory domain architecture at promoters and enhancers. *Nature Genetics*, 50(11):1533–1541, nov 2018. ISSN 1061-4036. doi: 10.1038/s41588-018-0234-5. URL <http://www.nature.com/articles/s41588-018-0234-5>.
- [239] M. Tomkova, M. McClellan, S. Kriaucionis, and B. Schuster-Boeckler. 5-hydroxymethylcytosine marks regions with reduced mutation frequency in human DNA. *eLife*, 5, may 2016. ISSN 2050-084X. doi: 10.7554/eLife.17082. URL <https://elifesciences.org/articles/17082>.
- [240] M. Toyota, N. Ahuja, M. Ohe-Toyota, J. G. Herman, S. B. Baylin, and J.-P. J. Issa. CpG island methylator phenotype in colorectal cancer. *Proceedings of the National Academy of Sciences*, 96(15):8681–8686, jul 1999. ISSN 0027-8424. doi: 10.1073/pnas.96.15.8681. URL <http://www.pnas.org/cgi/doi/10.1073/pnas.96.15.8681>.
- [241] A. Tsagaratou, T. Aijo, C.-W. J. Lio, X. Yue, Y. Huang, S. E. Jacobsen, H. Lahdesmaki, and A. Rao. Dissecting the dynamic changes of 5-hydroxymethylcytosine in T-cell development and differentiation. *Proceedings of the National Academy of Sciences*, 111(32):E3306–E3315, aug 2014. ISSN 0027-8424. doi: 10.1073/pnas.1412327111. URL <http://www.pnas.org/cgi/doi/10.1073/pnas.1412327111>.
- [242] B. Tycko. Allele-specific DNA methylation: beyond imprinting. *Human Molec-*

- ular Genetics*, 19(R2):R210–R220, oct 2010. ISSN 0964-6906. doi: 10.1093/hmg/ddq376. URL <https://academic.oup.com/hmg/article-lookup/doi/10.1093/hmg/ddq376>.
- [243] S. Uribe-Lewis, T. Carroll, S. Menon, A. Nicholson, P. J. Manasterski, D. J. Winton, S. J. A. Buczacki, and A. Murrell. 5-hydroxymethylcytosine and gene activity in mouse intestinal differentiation. *Scientific Reports*, 10(1):546, dec 2020. ISSN 2045-2322. doi: 10.1038/s41598-019-57214-z. URL <http://www.nature.com/articles/s41598-019-57214-z>.
- [244] A. Veloso, K. S. Kirkconnell, B. Magnuson, B. Biewen, M. T. Paulsen, T. E. Wilson, and M. Ljungman. Rate of elongation by RNA polymerase II is associated with specific gene features and epigenetic modifications. *Genome Research*, 24(6): 896–905, jun 2014. ISSN 1088-9051. doi: 10.1101/gr.171405.113. URL <http://genome.cshlp.org/cgi/doi/10.1101/gr.171405.113>.
- [245] A. Vihervaara, D. B. Mahat, M. J. Guertin, T. Chu, C. G. Danko, J. T. Lis, and L. Sistonen. Transcriptional response to stress is pre-wired by promoter and enhancer architecture. *Nature Communications*, 8(1):255, dec 2017. ISSN 2041-1723. doi: 10.1038/s41467-017-00151-0. URL <https://www.ncbi.nlm.nih.gov/pubmed/28811569><http://www.nature.com/articles/s41467-017-00151-0>.
- [246] A. Vihervaara, D. B. Mahat, S. Himanen, M. Blom, J. T. Lis, and L. Sistonen. Stress-Induced Transcriptional Memory Accelerates Promoter-Proximal Pause-Release and Decelerates Termination Over Mitotic Divisions. *SSRN Electronic Journal*, 2019. ISSN 1556-5068. doi: 10.2139/ssrn.3399580. URL <https://www.ssrn.com/abstract=3399580>.
- [247] J. Wan, V. F. Oliver, G. Wang, H. Zhu, D. J. Zack, S. L. Merbs, and J. Qian. Characterization of tissue-specific differential DNA methylation suggests distinct modes of positive and negative gene expression regulation. *BMC Genomics*, 16(1):49, dec 2015. ISSN 1471-2164. doi: 10.1186/s12864-015-1271-4. URL <https://bmcbgenomics.biomedcentral.com/articles/10.1186/s12864-015-1271-4>.

- [248] B. Wang, Z. Wang, D. Wang, B. Zhang, S.-G. Ong, M. Li, W. Yu, and Y. Wang. krCRISPR: an easy and efficient strategy for generating conditional knockout of essential genes in cells. *Journal of Biological Engineering*, 13(1): 35, dec 2019. ISSN 1754-1611. doi: 10.1186/s13036-019-0150-y. URL <https://jbioleng.biomedcentral.com/articles/10.1186/s13036-019-0150-y>.
- [249] L. Wang, Y. Zhou, L. Xu, R. Xiao, X. Lu, L. Chen, J. Chong, H. Li, C. He, X.-D. Fu, and D. Wang. Molecular basis for 5-carboxycytosine recognition by RNA polymerase II elongation complex. *Nature*, 523(7562):621–625, jul 2015. ISSN 0028-0836. doi: 10.1038/nature14482. URL <http://www.nature.com/articles/nature14482>.
- [250] W. Wang, L. Xu, L. Hu, J. Chong, C. He, and D. Wang. Epigenetic DNA Modification N 6 -Methyladenine Causes Site-Specific RNA Polymerase II Transcriptional Pausing. *Journal of the American Chemical Society*, 139(41):14436–14442, oct 2017. ISSN 0002-7863. doi: 10.1021/jacs.7b06381. URL <https://pubs.acs.org/doi/10.1021/jacs.7b06381>.
- [251] Z. Wang, M. Du, Q. Yuan, Y. Guo, J. N. Hutchinson, L. Su, Y. Zheng, J. Wang, L. A. Mucci, X. Lin, L. Hou, and D. C. Christiani. Epigenomic analysis of 5-hydroxymethylcytosine (5hmC) reveals novel DNA methylation markers for lung cancers. *Neoplasia*, 22(3):154–161, mar 2020. ISSN 14765586. doi: 10.1016/j.neo.2020.01.001. URL <https://linkinghub.elsevier.com/retrieve/pii/S1476558619303033>.
- [252] J. A. Watts, J. Burdick, J. Daigneault, Z. Zhu, C. Grunseich, A. Bruzel, and V. G. Cheung. cis Elements that Mediate RNA Polymerase II Pausing Regulate Human Gene Expression. *American Journal of Human Genetics*, 105(4):677–688, 2019. ISSN 15376605. doi: 10.1016/j.ajhg.2019.08.003. URL <https://doi.org/10.1016/j.ajhg.2019.08.003>.
- [253] A. R. Weber, C. Krawczyk, A. B. Robertson, A. Kuśnierczyk, C. B. Vågbø, D. Schuermann, A. Klungland, and P. Schär. Biochemical reconstitution of

- TET1–TDG–BER-dependent active DNA demethylation reveals a highly coordinated mechanism. *Nature Communications*, 7(1):10806, apr 2016. ISSN 2041-1723. doi: 10.1038/ncomms10806. URL <http://www.nature.com/articles/ncomms10806>.
- [254] M. Weber and D. Schübeler. Genomic patterns of DNA methylation: targets and function of an epigenetic mark. *Current Opinion in Cell Biology*, 19(3):273–280, jun 2007. ISSN 09550674. doi: 10.1016/j.ceb.2007.04.011. URL <https://linkinghub.elsevier.com/retrieve/pii/S0955067407000634>.
- [255] M. Weber, J. J. Davies, D. Wittig, E. J. Oakeley, M. Haase, W. L. Lam, and D. Schübeler. Chromosome-wide and promoter-specific analyses identify sites of differential DNA methylation in normal and transformed human cells. *Nature Genetics*, 37(8):853–862, aug 2005. ISSN 1061-4036. doi: 10.1038/ng1598. URL <http://www.nature.com/articles/ng1598>.
- [256] L. Wen, X. Li, L. Yan, Y. Tan, R. Li, Y. Zhao, Y. Wang, J. Xie, Y. Zhang, C. Song, M. Yu, X. Liu, P. Zhu, X. Li, Y. Hou, H. Guo, X. Wu, C. He, R. Li, F. Tang, and J. Qiao. Whole-genome analysis of 5-hydroxymethylcytosine and 5-methylcytosine at base resolution in the human brain. *Genome Biology*, 15(3):R49, 2014. ISSN 1465-6906. doi: 10.1186/gb-2014-15-3-r49. URL <http://genomebiology.biomedcentral.com/articles/10.1186/gb-2014-15-3-r49>.
- [257] K. Williams, J. Christensen, M. T. Pedersen, J. V. Johansen, P. A. C. Cloos, J. Rappsilber, and K. Helin. TET1 and hydroxymethylcytosine in transcription and DNA methylation fidelity. *Nature*, 473(7347):343–348, may 2011. ISSN 0028-0836. doi: 10.1038/nature10066. URL <http://www.nature.com/articles/nature10066>.
- [258] E. R. Wilson, N. M. Helton, S. E. Heath, R. S. Fulton, J. S. Welch, M. J. Walter, P. Westervelt, J. F. DiPersio, C. A. Miller, T. J. Ley, and D. H. Spencer. Genome-wide analysis of focal DNA hypermethylation in IDH-mutant AML samples 1 2. *bioRxiv*, 2021. doi: <https://doi.org/10.1101/2021.03.03.433799>.

- [259] E. M. Wissink, A. Vihervaara, N. D. Tipples, and J. T. Lis. Nascent RNA analyses: tracking transcription and its regulation. *Nature Reviews Genetics*, 20(12):705–723, dec 2019. ISSN 1471-0056. doi: 10.1038/s41576-019-0159-6. URL <http://www.nature.com/articles/s41576-019-0159-6>.
- [260] K. H. Wong, Y. Jin, and K. Struhl. TFIIF Phosphorylation of the Pol II CTD Stimulates Mediator Dissociation from the Preinitiation Complex and Promoter Escape. *Molecular Cell*, 54(4):601–612, may 2014. ISSN 10972765. doi: 10.1016/j.molcel.2014.03.024. URL <https://linkinghub.elsevier.com/retrieve/pii/S1097276514002573>.
- [261] Y. M. Woo, Y. Kwak, S. Namkoong, K. Kristjánsdóttir, S. H. Lee, J. H. Lee, and H. Kwak. TED-Seq Identifies the Dynamics of Poly(A) Length during ER Stress. *Cell Reports*, 24(13):3630–3641.e7, sep 2018. ISSN 22111247. doi: 10.1016/j.celrep.2018.08.084. URL <https://linkinghub.elsevier.com/retrieve/pii/S2211124718313883>.
- [262] H. Wu, A. C. D’Alessio, S. Ito, Z. Wang, K. Cui, K. Zhao, Y. E. Sun, and Y. Zhang. Genome-wide analysis of 5-hydroxymethylcytosine distribution reveals its dual function in transcriptional regulation in mouse embryonic stem cells. *Genes & Development*, 25(7):679–684, apr 2011. ISSN 0890-9369. doi: 10.1101/gad.2036011. URL <http://genesdev.cshlp.org/cgi/doi/10.1101/gad.2036011>.
- [263] G. R. Wyatt. Occurrence of 5-Methyl-Cytosine in Nucleic Acids. *Nature*, 166(4214):237–238, aug 1950. ISSN 0028-0836. doi: 10.1038/166237b0. URL <http://www.nature.com/articles/166237b0>.
- [264] G. R. Wyatt and S. S. Cohen. A New Pyrimidine Base from Bacteriophage Nucleic Acids. *Nature*, 170(4338):1072–1073, dec 1952. ISSN 0028-0836. doi: 10.1038/1701072a0. URL <http://www.nature.com/articles/1701072a0>.
- [265] Z. Xiao, W. Wu, C. Wu, M. Li, F. Sun, L. Zheng, G. Liu, X. Li, Z. Yun, J. Tang, Y. Yu, S. Luo, W. Sun, X. Feng, Q. Cheng, X. Tao, S. Wu, and J. Tao. 5-Hydroxymethylcytosine signature in circulating cell-free DNA as a po-

- tential diagnostic factor for early-stage colorectal cancer and precancerous adenoma. *Molecular Oncology*, 15(1):138–150, jan 2021. ISSN 1574-7891. doi: 10.1002/1878-0261.12833. URL <https://onlinelibrary.wiley.com/doi/10.1002/1878-0261.12833>.
- [266] X. Yang, H. Han, D. D. De Carvalho, F. D. Lay, P. A. Jones, and G. Liang. Gene Body Methylation Can Alter Gene Expression and Is a Therapeutic Target in Cancer. *Cancer Cell*, 26(4):577–590, oct 2014. ISSN 15356108. doi: 10.1016/j.ccr.2014.07.028. URL <https://linkinghub.elsevier.com/retrieve/pii/S153561081400316X>.
- [267] J. A. Yoder, N. S. Soman, G. L. Verdine, and T. H. Bestor. DNA (cytosine-5)-methyltransferases in mouse cells and tissues. studies with a mechanism-based probe. *Journal of Molecular Biology*, 270(3):385–395, jul 1997. ISSN 00222836. doi: 10.1006/jmbi.1997.1125. URL <https://linkinghub.elsevier.com/retrieve/pii/S0022283697911256>.
- [268] M. Yu, G. C. Hon, K. E. Szulwach, C.-X. Song, L. Zhang, A. Kim, X. Li, Q. Dai, Y. Shen, B. Park, J.-H. Min, P. Jin, B. Ren, and C. He. Base-Resolution Analysis of 5-Hydroxymethylcytosine in the Mammalian Genome. *Cell*, 149(6):1368–1380, jun 2012. ISSN 00928674. doi: 10.1016/j.cell.2012.04.027. URL <https://linkinghub.elsevier.com/retrieve/pii/S009286741200534X>.
- [269] A. Zemach, I. E. McDaniel, P. Silva, and D. Zilberman. Genome-Wide Evolutionary Analysis of Eukaryotic DNA Methylation. *Science*, 328(5980):916–919, may 2010. ISSN 0036-8075. doi: 10.1126/science.1186366. URL <https://www.sciencemag.org/lookup/doi/10.1126/science.1186366>.
- [270] J. Zhang, M. Cavallaro, and D. Hebenstreit. Timing Polymerase Pausing with TV-PRO-seq. *bioRxiv*, 2018. doi: bioRxiv:10.1101/461442. URL <https://www.biorxiv.org/content/early/2018/11/05/461442>.
- [271] L. Zhang, X. Lu, J. Lu, H. Liang, Q. Dai, G.-L. Xu, C. Luo, H. Jiang, and C. He. Thymine DNA glycosylase specifically recognizes 5-carboxylcytosine-

modified DNA. *Nature Chemical Biology*, 8(4):328–330, apr 2012. ISSN 1552-4450. doi: 10.1038/nchembio.914. URL <http://www.nature.com/articles/nchembio.914>.

- [272] Y. Zhang, K. Wu, Y. Shao, F. Sui, Q. Yang, B. Shi, P. Hou, and M. Ji. Decreased 5-Hydroxymethylcytosine (5-hmC) predicts poor prognosis in early-stage laryngeal squamous cell carcinoma. *American Journal of Cancer Research*, 6(5), 2016. ISSN 21566976.
- [273] L. Zhao, M.-a. Sun, Z. Li, X. Bai, M. Yu, M. Wang, L. Liang, X. Shao, S. Arnovitz, Q. Wang, C. He, X. Lu, J. Chen, and H. Xie. The dynamics of DNA methylation fidelity during mouse embryonic stem cell self-renewal and differentiation. *Genome Research*, 24(8):1296–1307, aug 2014. ISSN 1088-9051. doi: 10.1101/gr.163147.113. URL <http://genome.cshlp.org/lookup/doi/10.1101/gr.163147.113>.

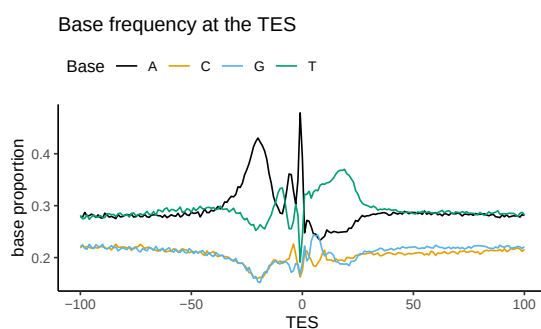


Figure S1: **Base distribution at annotated TES** Base frequency upstream and downstream of annotated TESs.