Genome wide efficiency profiling reveals modulation of maintenance and de novo methylation by Tets

Pascal Giehr *[1], Charalampos Kyriakopoulos *[2], Karl Nordström[1], Abduhlrahman Salhab[1], Fabian Müller[5], Ferdinand von Meyenn[6], Gabriella Ficz[7], Wolf Reik[8], Verena Wolf †[2], and Jörn Walter †[1]

[1]Department of Genetics and Epigenetics, Saarland University, Campus A2.4,66123 Saarbrücken, Germany

[2]Computer Science Department, Saarland University, Campus E1.3, 66123 Saarbrücken, Germany

[5]Genetics Department, Stanford University, 94305 Stanford, USA

[6]Department of Health Sciences and Technologys, ETH Zürich, Schorenstrasse 16, 8603 Schwerzenbach, Switzerland

[7]Haemato-Oncology, Queen Mary University of London, EC1M6BQ London, United Kingdom

[8]Epigenetics Department, Babraham Institute, CB223AT Cambridge, United Kingdom

Abstract

Background DNA methylation is an essential epigenetic modification which is set and maintained by DNA methyl transferases (Dnmts) and removed via active and passive mechanisms involving Tet mediated oxidation. While the molecular mechanisms of these enzymes are well studied, their interplay on shaping cell specific methylomes remains less well understood. In our work we model the activities of Tets and Dnmts at single CpGs across the genome using a novel type of high resolution sequencing data.

Results To accurately measure 5mC and 5hmC levels at single CpGs we developed RRHPoxBS, a reduced representation hairpin oxidative bisulfite sequencing approach. Using this method we mapped the methylomes and hydroxymethylomes of wild type and Tet triple knockout mouse embryonic stem cells. These comprehensive datasets were then used to develop an extended Hidden Markov model allowing us i) to determine the symmetrical methylation and hydroxymethylation state at millions of individual CpGs, ii) infer the maintenance and *de novo* methylation efficiencies of Dnmts and the hydroxylation efficiencies of Tets at individual CpG positions. We find that Tets exhibit their highest activity around unmethylated regulatory elements, i.e. active promoters and enhancers. Furthermore, we find that Tets' presence has a profound effect on the global and local maintenance and *de novo* methylation activities by the Dnmts, not only substantially contributing to a universal demethylation of the genome but also shaping the overall methylation landscape.

Conclusions Our analysis demonstrates that a fine tuned and locally controlled interplay between Tets and Dnmts is important to modulate *de novo* and maintenance

*These authors contributed equally to this work
†Correspondence: wolf@cs.uni-saarland.de (VW), j.walter@mx.uni-saarland.de (JW)

activities of Dnmts across the genome. Tet activities contribute to DNA methylation patterning in the following ways: They oxidize 5mC, they locally shield DNA from accidental *de novo* methylation and at the same time modulate maintenance and *de novo* methylation efficiencies of Dnmts across the genome.

Background

Genetic information encoded in the DNA is regulated by epigenetic mechanisms, such as DNA methylation [1, 2, 3, 4]. In mammals, the methylation of DNA is restricted to cytosine and is almost exclusively found in a palindromic CpG dinucleotide context [5, 6, 7]. Generation of 5-methylcytosine (5mC) is catalyzed by the DNA methyltransferases Dnmt1, Dnmt3a, Dnmt3b and Dnmt3c. These enzymes catalyze the transfer of a methyl group from s-adenosyl methionine to the fifth carbon atom of cytosine.

Even though under certain conditions Dnmt1 has been shown to methylate also unmethylated CpGs [16, 17, 18], this enzyme is mainly responsible for maintaining existing methylation patterns after replication. Via interaction with Uhrf1 and PCNA, Dnmt1 is tightly associated with the replication machinery [8, 9]. Furthermore, the cooperation with Uhrf1 modulates Dnmt1 to be receptive for hemimethylated DNA generated after replication [10, 11]. Thus, the protein complex postreplicatively copies the methylation pattern from the inherited to the newly synthesized DNA strand [12, 13].

Dnmt3a, Dnmt3b and Dnmt3c methylate DNA independently of its methylation status (hemimethylated or unmethylated) and are mainly responsible for the establishment of new methylation patterns during development [14, 15].

Once established, 5mC can be further processed by a family of di-oxigenases, the ten-eleven translocation enzymes Tet1, Tet2 and Tet3 [19, 20, 21]. These Fe(II) and oxo-glutarate-dependent enzymes consecutively oxidize 5mC to 5-hydroxymethyl cy-

tosine (5hmC), 5-formyl cytosine (5fC) and ultimately to 5-carboxy cytosine (5caC) [22, 23]. 5hmC is the most abundant oxidative variant and can be found in numerous cell types [24, 25, 26]. Each oxidation step changes the chemical properties of the base and with it its biological function [27, 28, 29]. Several mechanisms have been proposed in which oxidative cytosine derivatives (oxC) serve as an intermediate during the course of active or passive demethylation [30, 31, 32, 33, 34]. Such removal of 5mC occurs locally during cell differentiation, but also on a genome-wide scale in the zygote, as well as during the maturation of primordial germ cells (PGCs) [35, 36, 37]. Global loss of 5mC has been observed in cultivated mouse embryonic stem cells (ES cells) during their transition from Serum to 2i medium (2i). Under classical Serum/LIF conditions, ES cells exhibit DNA hypermethylation, whereas upon transition to GSK3β and Erk1/2 inhibitors (2i containing medium), the cells experience a gradual genome-wide loss of 5mC [38, 39, 40]. Even though the enzymatic mechanisms of oxCs generation are well characterized, the question how oxCs are inherited across replication, as well as the impact of Tets and oxCs on maintaining or changing an existing methylome remains elusive.

To address these questions we developed and applied **R**educed **R**epresentation **H**air**p**in **ox**idative **B**isulfite **S**equencing (RRHPoxBS), which combines three features: (i) a genome-wide detection of a representative number of CpGs (RRBS), (ii) a strand-specific detection of 5mC by using a short hairpin oligo (HPBS) and lastly, (iii) the localization of 5mC and 5hmC, respectively, by combining regular and oxidative bisulfite sequencing (oxBS) [41, 42, 43, 44, 45]. We use then these deep RRHoxBS data as input for an extended version of a Hidden Markov Model (HMM) first presented in [46, 47], to predict the levels and the strand specific distribution of 5mC and 5hmC and estimate the enzymatic activities of Dnmts for maintenance methylation, *de novo* methylation, and hydroxylation of Tets in 2i over time. Finally we interpret these data in the context of the genome and other epigenetic features.

We find a very specific spatial distribution of Tet activity and show that DNA methylation and hydroxylation efficiencies of Dnmts and Tets are negatively correlated throughout the genome in a very consistent fashion. Finally, we infer that the absence of Tet enzymes in Tet triple knockout (TKO) cells changes maintenance and *de novo* efficiency profiles in regions, which are protected by Tet enzymes in wild type (WT) ES cells.

Methods

Reduced Representation Hairpin oxidative Bisulfite Sequencing (RRHPoxBS)

1.2μg DNA was divided equally into three 0.5ml reaction tubes and digested in a 20μl reaction using 20U of HaeIII (NEB), AluI (NEB) and 10U HpyCH4V (NEB), respectively. The reactions were incubated overnight for a minimum of 12h at 37°C. Restriction enzymes were inactivated at 80°C for 30min. The reactions were pooled and purified using 2x AmpureXP beads (120μl) from Beckman Coulter. DNA was eluted in 18μl ddH_2O and subjected to A-tailing by adding 1μl dATP(1mM) and 1μl Klenow exo-(5U/μl, NEB), incubated at 37°C followed by an inactivation at 75°C for 30min. Next, hairpin linker and sequencing adapter were ligated to opposed ends of each restriction fragment. For this, 1μl biotin labeled hairpin linker (100μM), 0.5μl sequencing adapter(100μM), 2.5μl ATP(10mM), as well as 1μl T4 DNA Ligase (200U/μl, NEB) were added to the A-tailed DNA and incubated for 16h at 16°C. The reaction was purified using AMPureXP(2x) beads followed by enrichment for hairpin containing fragments with streptavidin beads (Dynabeads™M-280 Streptavidin, ThermoFischer). The library was then subjected to BS/oxBS work-flow of the TrueMethyl kit from CEGX according to manufacturer's instructions. Amplification of the library was done using HotStarTaq® polymerase (Qiagen) and sequencing

was performed on an Illumina HiSeq2500 platform in a 150bp paired-end sequencing mode. We generated six hairpin libraries (BS and oxBS, respectively) for WT ES cells at three distinct time points - Serum/Lif(d0), 2i 72h (d3) and 2i 144h (d6), as well as four libraries for TetTKO cells (BS only) - Serum/LIF, 2i 48h (d2), 2i 96h (d4) and 168h (d7).

Read Mapping and Methylation Calling

The sequences were aligned as suggested by Porter *et al.* [48]. In detail; reads were trimmed for adapter, hairpin-linker and 3' quality (Q$\geq$20) with TrimGalore! [49] and cutadapt [50]. Trimmed read pairs were aligned with the Smith-Waterman algorithm allowing for bisulfite induced mismatches. The two bisulfite converted strands were used to deduce the original genomic sequence. Mismatches other than G-to-A and T-to-C were replaced with N. The resulting sequences were aligned to the mouse genome (mm10) with GEM-mapper (beta build 1.376) [51], after which the methylation information was reintroduced with a custom pileup function based on HTSJDK [52] and ratios for the four methylation states were calculated for each cytosine. The pipeline was implemented with the Ruffus pipeline framework [53].

Statistical Modeling

Estimation of (hydroxy-)methylation levels

For CpGs with observations at up to two time points we combined information from BS and oxBS experiments to arrive at maximum likelihood estimates (MLEs) for strand specific (hydroxy-)methylation levels for each observation time point. The derived MLEs take into account the conversion errors of each experiment and we estimate their accuracy by approximating the corresponding standard deviations. For details see SI Sec. 1 and Sec. 2.

Estimation of Enzymatic Efficiencies

For CpGs for which the maximal number of three observation time points is available, we defined an underlying discrete time Markov process that shapes the demethylation dynamics. The state space of the process is the set of possible CpG's site state $s \in \mathcal{S} = \{u, m, h\}^2$, where state s encodes whether the upper and the lower strand of the site is *unmethylated* (u), *methylated* (m) or *hydroxylated* (h). E.g. in state (h, u) the upper strand is hydroxylated and the lower strand is unmethylated. The model's time parameter corresponds to the number of cell divisions and the transitions of a state are being triggered by consecutive division or (hydroxy-) methylation events. Getting measurements along with the conversion errors from two different experiments (BS and oxBS) allows us to define one HMM for each experiment and get accurate estimates for the model's parameters. The latter are linear functions that represent the enzymes' efficiencies over time. In addition, a parameter related to the recognition of 5hmC by Dnmts (passive demethylation) is being estimated for each CpG. For a detailed presentation of the above model we refer the reader to Giehr *et al.* 2016 [47, 46] and to SI Sec. 1 and 2.

In case of an adequate number of samples per time point when a very deep sequencing is possible, the MLE provides accurate estimates with narrow confidence intervals [47]. On the other hand, MLE is known to give imprecise results for a smaller number of samples [54, 55] and in particular in cases where the true values are close to the boundary constraints [56]. Since a consistently deep sequencing ($\geq$100x) is under the current methods impossible on a genome-wide level, we develop here a combination of MLE and Bayesian Inference (BI) methods in order to get accurate estimates even in case of a lower coverage, common for genome wide applications. In particular, we use a MLE step as initial information to be given to a Metropolis-Hastings MCMC sampler from which we get the posterior distribution of the parameters. The approach is being described in detail in SI Sec. 3.

Clustering of Single CpG Efficiencies

To identify CpGs with similar enzymatic profiles we cluster the genome-wide output of our model, meaning the efficiencies of the enzymes responsible for maintenance methylation, *de novo* methylation and hydroxylation over time for $1.5 \cdot 10^6$ CpGs uniformly located across the entire genome. Since we aim to cluster parameter estimates we consider k-error, a sophisticated clustering approach that takes into account the uncertainty, i.e., posterior's covariance matrix, around the BI estimators, i.e., posterior's mean. The clustering approach we apply here returns different and probably more "natural" clusters than a typical k-means clustering algorithm would return. We determine the optimal number of clusters based on two different criteria, Davies-Bouldin Criterion and the elbow method. For details of the individual clustering approaches we refer to SI Sec. 3.1.

Segmentation

The whole genome bisulfite data of primed mouse ES cells (Ficz *et al.* 2013) was segmented into low methylation regions (LMRs), unmethylated regions (UMRs) and partially methylated domains (PMDs) [38], using MethylSeekR [57]. The rest of the genome, after filtering gaps annotated by UCSC, was called highly methylated regions (HMRs) [58]. The aggregated strand information per CpG was used as an input for MethylSeekR. The applied parameters were the following; coverage of $\geq$5x, $\leq$50% methylation and FDR $<$0.05 for calling hypomethylation regions and consequently a cutoff of $\geq$ 4 CpGs per LMR.

LOLA Analysis

We performed a standard LOLA analysis against the regular LOLA universe, extended by ChIP-Seq profiles from von Meyenn *et al.* 2016 (GSE70724, GSE77420), Walter *et al.* 2016 (GSE71593) and Encode. [40, 59, 60].

Chip-Seq Data Processing

Low quality read tails (Q¡20) and adapter sequences were trimmed using TrimGalore! (version 0.4.2) (`http://www.bioinformatics.babraham.ac.uk/projects/trim_galore`). Trimmed reads were aligned to the mouse reference genome (mm10) using GEM mapper (version 1.376 beta) (Marco-Sola *et al.*, 2012). Samtools (version 1.3) (Li *et al.*, 2009) was used to convert SAM to BAM format. MarkDuplicate (version 1.115) from Picard tools (`http://broadinstitute.github.io/picard/`) was used to mark the PCR duplications. Normalized coverage files with respect to library size were generated using deepTools v1.5.9.1 (Ramírez *et al.*, 2014) with bamCoverage command.

Hi-C Data Processing

Homer tool was used to process Hi-C data (PMID: 20513432, PMID:30146161). Reads were trimmed considering DpnII (GATC) as a restriction enzyme in the Hi-C assay. Mates were aligned separately to mosue genome (mm10) using bowtie2 (`https://doi.org/10.1038/nmeth.1923`). PCR duplicates were removed. A tag directory was generated with makeTagDirectory command which was used afterwards by runHiCpca.pl command to identify A/B compartments at 25kb resolution and 50kb window (super resolution). Positive values were assigned to A compartments while the negative ones were assigned to B compartments.

Results

In our study we used a well established ES cell system to precisely map 5mC and 5hmC across the genome in a time series of experiments and to study the enzymatic contribution of Tets and Dnmts for the progressive genome wide DNA (de)methylation. For this we first needed to generate a high resolution data set

based on a novel genome wide hairpin sequencing approach, **R**educed **R**epresentation **H**airpin **ox**idative **B**isulfite **S**equencing (RRHPoxBS). The design of the RRHPoxBS approach covered up to 4 million CpG dyads for which we could infer the precise distributions of 5mC and 5hmC. Following a strict read and conversion quality control, we filtered for sufficient sequencing depth and ended up with about 2 million CpGs per sample for subsequent comparative modeling. To follow the dynamics of the enzymes over time we generated six data sets for WT ES cells, i.e. BS and oxBS libraries for three different time points, starting with Serum/Lif (d0), followed by 72h 2i (d3) and 144h 2i (d6). For a comparison we also generated four datasets for Tet TKO cells starting with Serum/Lif (d0) followed by 48h in 2i (d2), 96h in 2i (d4) and 168h in 2i (d7).

Impaired loss of 5mC in Tet TKO ES cells

Under primed conditions (Serum/LIF) WT ES cells show a overall level of 78% methylation. About 56% of CpGs are fully methylated while $\approx 22\%$ are found in a hemimethylated state (Fig.: 1a). Among cultivation in 2i medium the DNA becomes progressively demethylated, such that after 6 days in 2i only 30% of CpGs retain a methylated state (fully or hemimethylated). For all time points, we find that hemimethylation is equally distributed among both DNA strands (Fig.: 1a and SI Sec. 4.3 Figure 23).

In addition, we observe that oxBS samples always display lower methylation levels than BS samples (Fig.: 1a). This difference corresponds to the amount of 5hmC of each sample. We detect the main difference in the hemimethylated proportion, indicating that a considerable amount of 5hmC might exist in a hemi(hydroxy)methylated (5hmC/C or C/5hmC) state (Fig.: 1a). Initially, the amount of 5hmC is quite low but we observe a notable increase at d3, while later (d6) the amount of 5hmC decreases again (Fig.: 1a).

ES cells lacking Tet enzymes (Tet TKO) show a marginal increase of methylated CpG dyads (82% fully- or hemimethylated, Fig.: 1c) in comparison to WT under primed conditions (Serum/LIF). However, in relation to WT the TKO cells show a higher frequency of fully methylated CpGs ($\approx 72\%$) and a reduced proportion of hemimethylated CpGs (hemiCpGs; $\approx 10\%$) (Fig.: 1c). We concluded that in WT ES cells, the enhanced presence of hemiCpGs is directly coupled to 5mC oxidation by Tets.

On a first glimpse Tet TKO ES cells might show similar kinetics of DNA demethylation as WT ES cells (Fig.: 1c). However, in contrast to previous data [60] the RRHoxBS data allow us to precisely estimate the demethylation kinetics revealing that in WT ES cells the generation of unmethylated cytosine is 8% per day, while in Tet TKO cells it drops to 4.2% (SI Sec. 4.4 Figure 24). This indicates that the presence of Tets has a considerable influence on DNA demethylation kinetics.

RRHPoxBS sequencing also allowed us to accurately determine the amount, location and distribution of non-CpG methylation in (WT and Tet TKO) ES cells. In both, WT and TKO, we find CpA to be the most frequent methylated non-CpG motif (SI Sec. 4.7 Figure 32 and 33). Over time, non-CpG methylation becomes gradually reduced upon cultivation in 2i. In WT ES cells the number of methylated non-CpGs was identical in BS and oxBS libraries, indicating that non-CpGs are not a substrate for Tet oxidation (Fig.: 1b). In Tet TKO cells, the number of methylated non-CpGs is approximately doubled as compared to WT ES cells. Since non-CpG methylation is strictly dependent on the presence of *de novo* methylation activities by Dnmt3a/b, the higher non-CpG methylation in TKO cells, both under primed ($\geqq 2\%$) and naive ($\geqq 0.6\%$ after 168h 2i) conditions (Fig.: 1d), points clearly towards an increased *de novo* methylation activity by Dnmt3a/b in the absence of Tet enzymes.

Tet TKO ES cells retain de novo methylation activity

The genome wide observations in WT and TKO cells prompted us to deeper examine the relative contributions of Dnmts and Tets at individual CpGs. The developed HMM approach considers conversion errors of BS and oxBS treatment and estimates accurate 5mC and 5hmC levels, as well as the strand specific distribution of both cytosine forms. Furthermore, combining the strengths of MLE and BI our model estimates at every individual CpG the enzyme efficiencies of $i)$ maintenance methylation, $ii)$ *de novo* methylation and $iii)$ hydroxylation enzymes that explain the observed dynamics of the (hydroxy-)methylation patterns over time.

The estimated methylation levels for WT and Tet TKO ES cells fit nicely to the hairpin methylation data (SI Sec. 3, Fig.: 3), indicating a high accuracy of our model. Consequently, we observe a constant decline of fully methylated CpGs in WT ES cells over time (Fig.: 2a). Moreover, in WT ES cells, the HMM estimates a notable amount of 5hmC at all time points. Note that the displayed amount of 5hmC (yellow) refers to the sum of all possible 5hmC states, i.e., 5hmC/5hmC, 5hmC/5mC, 5mC/5hmC, 5hmC/C, C/5hmC. The highest amount of 5hmC is observed at d3, meaning that WT ES cells display a transient increase of 5hmC after cultivation in 2i. We observe a similar behaviour for hemiCpGs in WT ES cells. The parameter estimation by our model illustrates a mean maintenance methylation efficiency of about 61.4% at d0, which remains almost constant over time (60.1% at d6) (Fig.: 2b). In contrast, *de novo* methylation efficiency shows a strong decrease (from 14.1% to 4.5% at d6) and the hydroxylation efficiency an increase (from 22.2% at d0 to 29.1% at d6) over time. This observation is in agreement with previous observations which demonstrated a reduction in RNA and protein levels of Dnmt3a/b in 2i, but an increased expression of Tet1/2 on a genome wide level [38, 60].

In Tet TKO cells maintenance efficiency lies by 58.8% at d0, which represents

a marginal reduction compared to WT ES cells. Similarly to WT, maintenance efficiency remains stable over time (58.6% at d7) also in Tet TKO cells.

The most pronounced difference between WT and Tet TKO cells we see in *de novo* methylation efficiency. More specifically in Tet TKO *de novo* begins from (20.3%) at day 0 and exhibits only a slight decrease over time (16.2% at d7).

Finally, the model output also confirms the observed by the hairpin methylation data reduced demethylation rate of Tet TKO cells and suggests a substantial contribution of 5hmC and Tet enzyme on DNA demethylation. In fact, the model favors a scenario in which 5hmC is less well recognized (probability of non recognition equals $p = 0.66$, SI Sec. 1.1) by the maintenance machinery after replication, promoting a faster demethylation process.

Tets prevent the spreading of DNA methylation

We next related the model estimates to genomic, enzymatic and epigenetic features first focusing on Dnmt and Tet enzyme efficiencies across large genome segments with distinct methylation states.

We used MethylSeekR [57], to partition the genome into four states: highly methylated regions (HMRs), partially methylated domains (PMDs), low methylated regions (LMRs) and unmethylated regions (UMRs). The segmentation was performed on whole genome bisulfite sequencing (WGBS) data from WT ES cells cultivated under Serum/Lif conditions (Ficz *et al.* 2013) on the identical cell batch used for our study [38]. The estimated methylation levels (sum of 5mC and 5hmC) for WT ES cells by our model fully agreed with those derived from WGBS (SI Sec. 4.8, Fig.: 35.C and 35.E). This not only confirmed the accuracy of our model output but also denoted that we can use the precise WGBS segmentation for further analysis. We found that the majority of the WT ES cell genome (86%) consists of large HMRs (SI Sec. 4.8, Fig.: 35.A and Fig.: 35.B) followed by shorter ($\approx 13\%$)

PMDs. The residual 2% of the genome are found to be LMRs (0.4%) or UMRs (1.5%)(SI Sec. 4.8, Fig.: 35.A and Fig.: 35.B).

Next we assigned the 5mC and 5hmC modification levels, their distribution at CpG dyads and the corresponding Dnmt/Tet enzyme efficiencies determined by our model to CpGs in the individual segments (Fig.: 3a and 3b). In WT ES cells, all segments show a continuous reduction in DNA methylation over time. This is particularly evident in segments with initial high 5mC levels, i.e., HMRs and PMDs. HMRs and PMDs also exhibit the highest amount of 5hmC and hemiCpGs which transiently increases at d3 and d4 (Fig.: 3a). LMRs and UMRs show different kinetics as both the amounts of 5hmC and hemiCpGs constantly decline over time. The increase of hemiCpGs in HMRs and PMDs is a clear sign of impaired maintenance methylation in naive ES cells linked to the reported temporal increase in Tet expression and loss of Dnmt1 activity [38, 60].

Based on the methylation data, our model predicts high maintenance methylation efficiency in HMRs ($\approx 69\%$) and PMDs ($\approx 61\%$), but low maintenance efficiency in LMRs ($\approx 32\%$) and UMRs ($\approx 26\%$). Additionally, we observe a relatively high *de novo* methylation efficiency at HMRs ($\approx 18\%$) in primed ES cells. Overall, *de novo* methylation efficiency strongly decreases upon cultivation in 2i, which corresponds well with the previous described loss of Dnmt3a/b under these conditions. In contrast, hydroxylation efficiency is high in UMRs ($\approx 63\%$) and LMRs ($\approx 55\%$), but low in HMRs ($\approx 13\%$) and PMDs ($\approx 24\%$). Together, our results indicate regional differences and an antagonistic behaviour of Dnmts and Tets. This antagonism has been validated by estimating a robust spatial correlation measure between the efficiencies of Dnmts and Tets across the whole genome (SI Sec. 3.3, Fig.: 12).

Both WT and TKO ES cells show overall a decline of DNA methylation across all segments over time (Fig.:3a). However, in TKO cells, all segments retain a

substantial higher frequency of fully methylated CpG dyads across all time points. This observation indicates a reduced demethylation rate in all segments under the absence of Tets. Surprisingly, under primed conditions (d0) Tet TKO cells show a higher number of unmethylated CpGs in HMRs as compared to WT ES cells.

Most importantly, comparing the Tet TKO with WT ES cells, we observe a strong change in Dnmt efficiencies. Maintenance methylation efficiency shows a reduction in HMRs and PMDs of TKO cells, while it clearly increases in LMRs and UMRs (Fig.: 3b), resulting in almost equal maintenance activity across all segments. In the case of *de novo* methylation efficiency, we observe a more stable and slightly increased activity in all segments.

Tet efficiency marks active and poised regulatory regions

To deeper dwell into the spatial distribution of the various enzymatic profiles accross the genome we performed a clustering of all CpGs based on their individual enzymatic efficiencies (along with their temporal changes) in WT ES cells. Following our approach (SI Sec. 3.1, 3.2) we identified three clusters with distinct profiles of enzymatic activity (Fig.: 4).

Cluster 1 comprises 255492 CpGs (SI Sec. 4.2, Fig.: 37.A) which are characterized by low methylation levels (5mC of 14%, 5hmC of 11% at d0) (Fig.: 4b). Accordingly, these CpGs are located mainly in UMRs and, to a lesser extent, in LMRs (SI Sec. 4.8, Fig.: 37.B). CpGs of Cluster 1 exhibit relatively low maintenance methylation (40% at d0) and almost no *de novo* activity. Over time, maintenance methylation slightly increases to 45% at d6. In addition, these CpGs display a very high hydroxylation efficiency which remains stable over time (66% at d0 to 64% at d6) (Fig.: 4a).

Cluster 2 contains 202562 CpGs (SI Sec. 4.8, Fig.: 37.A) and initially displays a high methylation level (60% fully methylated CpGs). In addition, CpGs of Cluster

2 also show a considerable amount of 5hmC (21%). While the level of 5mC rapidly declines upon transition to 2i, 5hmC displays a mild transient increase at d3 (24%) (Fig.: 4b). CpGs of Cluster 2 are dispersed across HMRs and PMDs (SI Sec. 4.8, Fig.: 37.B). At d0 CpGs assigned to Cluster 2 display a high *de novo* methylation efficiency (20%) but relatively low maintenance methylation (27%), as well as an average hydroxylation efficiency (20%). Upon cultivation in 2i, *de novo* methylation efficiency continuously declines to almost zero percent at d6. In contrast, both maintenance methylation and hydroxylation efficiency display a notable increase over time (Fig.: 4a).

Cluster 3 comprises 1073476 CpGs (SI Sec. 4.8, Fig.: 37.A) and among the three clusters, displays the highest amount of 5mC and 5hmC (64% fully methylated CpGs, 20% 5hmC at d0) (Fig.: 4b). Similar to Cluster 2, we observe a constant decrease of 5mC, while 5hmC shows a transient increase at d3 ($> 26\%$). CpGs of Cluster 3 exhibit both high maintenance (69%), as well as high *de novo* methylation efficiency (25%) at d0. Overtime, we observe a mild reduction in maintenance (60% at d6) and strong reduction of *de novo* methylation efficiency (4% at d6). In addition, Cluster 3 exhibits a low hydroxylation efficiency (14% at d0), which slightly increases over time (20% at d6) (Fig.: 4a). CpGs of Cluster 3 are mainly located in HMRs and PMDs but also appear frequently in UMRs (SI Sec. 4.8, Fig.: 37.B).

An enrichment analysis for CpGs of all three clusters using the LOLA package [59] provided a deeper insight into their association with genomic and epigenetic features (Fig.: 4c). CpGs of Cluster 1 were found to be enriched in regions with clear regulatory signatures. This included a broad enrichment for euchromatic histone marks (H3K4me3, H3K9ac, H3K27ac), binding sites for epigenetic modifiers (Ez2H, Kdm2a, Kdm2b, Yy1) and Tet1, general and specific transcriptional regulators (Myc, Sin3A, Tbp, Taf1, Taf3, Polr2a), as well as stem cell markers (Pou5f1,

Sox2, Myc and Nanog) (Fig.: 4c). In addition, we find a strong overlap of CpGs from Cluster 1 with UMRs, CpG islands and low complexity repeats. Thus, the enrichment profile of Cluster 1 indicates that Tet enzymes are more active at open and accessible chromatin.

CpGs of Cluster 2 and 3 are mainly located within repetitive elements (SINEs, LINEs, LTRs etc.), HMRS and PMDs, as well as domains with broad heterochromatic marks such as H3K9me2/3 (Fig.: 4c). However, while Cluster 3 displays exclusively heterochromatic signatures, CpGs of Cluster 2 are also partially located at transcription factor binding sites (TFBS), as well as domains containing bivalent and euchromatic histone marks such as H3K4me3 and H3K4me1, respectively (Fig.: 4c).

In Tet TKO ES cells, we observe considerable changes in the enzymatic efficiencies across all three clusters which are accompanied by changes in their methylation patterns. CpGs in Cluster 1 exhibit a notable increase in maintenance- (50% at d0) and to a smaller extent *de novo* methylation efficiency (5% at d0) which results in a higher frequency of fully methylated CpGs (27% at d0) and hemimethylated CpGs at later time points (12% at d4 and d6) (Fig.: 4a). Cluster 2 also shows a clear increase in maintenance methylation efficiency (d0: 59%), but at the same time a mild reduction in *de novo* methylation efficiency (d0: 18%). Moreover, we observe an increase in fully- and hemimethylated CpGs for all time points (from 75% at d0 to 35% day7). However, at d0, Tet TKO cells display a higher frequency of unmethylated CpGs (15%) in Cluster 2 compared to WT ES cells (Fig.: 4b). Cluster 3 exhibits a reduction in both maintenance- (60%) and *de novo* methylation efficiency (22%) in Tet TKO cells. Similar to Cluster 2 we observe a higher frequency of fully- and hemimethylated CpGs across all time points, but at the same time a higher frequency of unmethylated CpGs at d0 (12%). Overall, comparing the enzymatic activity between WT and Tet TKO cells we end up with similar observations as in

segments comparison. Under the absence of Tets Dnmt1 spreads uniformly across all clusters, while Dnmt3a/b activity remains stable over time.

Tets regulate Dnmts at TSS and TFBS

The genome wide antagonistic effects of Dnmts' and Tets' activity across segments and clusters prompted us to plot the enzymatic efficiencies of CpGs across genes, histone marks and ChIP profiled TFBS using DeepTools [62] (Fig.: 5) in order to investigate regularities and general local dependencies. In WT cells the enzymes' efficiencies across genes and TFBS show once more an opposing behavior: At transcription start sites (TSS) and TFBS, high hydroxylation efficiency is coupled to reduced methylation (both maintenance and *de novo*) efficiency. This inverse behavior at TSS remains upon 2i cultivation. *De novo* methylation almost disappears across the entire gene including the gene body. Under primed conditions *de novo* methylation is absent in TSS but has a strong presence in the gene body and it almost disappears from the entire gene over time after the transition to 2i. The observed efficiency profiles for maintenance methylation, *de novo* methylation and hydroxylation, correspond nicely to Uhrf1, Dnmt3a/b, as well as Tet1 ChIP profiles, respectively (SI Sec. 4.6 Fig. 30 and 31).

In Tet TKO ES cells the TSS associated drop in maintenance methylation is much less pronounced and almost absent at d6/d7. In addition, *de novo* methylation is only mildly reduced upon cultivation in 2i and clearly maintained across the gene body (Fig.: 5a). Regulatory regions marked by Sox2, H3K4me3 and Tet1 enrichment show a strong hydroxylation activity in WT cells which is inversely linked to an impaired maintenance and *de novo* methylation activity. Interestingly, the lack of Tet activity in TKO cells does not change *de novo* methylation but maintenance activity across regulatory regions (Fig.: 5b).

Discussion

In our study, we provide a comprehensive genome wide DNA-methylation modeling approach that allowed us to infer how the activity of Dnmts and Tets contribute to modify CpGs and non-CpGs across the genome in a functional context. This approach was only possible by applying a novel sequencing method generating high resolution methylome data at a single CpG resolution and in double stranded DNA. Our RRHPoxBS data are well in line with previous described overall methylation levels of mouse ES cells determined by classical RRBS or WGBS [38, 60]. In addition, RRHPoxBS data comprise three important new features: (i) a genome-wide representation of up to 4 million CpGs uniformly distributed across the genome, (ii) a precise determination of 5mC and 5hmC levels at a single CpG dyad and (iii) a precise mapping of hemimethylated states and positions of non-CpG methylation.

The overall evaluation of our RRHPoxBS data showed that hemiCpGs are almost equally distributed on both DNA strands following the behavior of symmetric CpG methylation. This suggests that hemimethylation is most likely the result of (strand-) undirected *de novo* methylation or active and passive demethylation events, respectively. Furthermore, we detect more hemimethylation in WT compared to Tet TKO cells, which indicates that Tet enzymes enhance the passive loss of 5mC. Indeed, our model predicts that 5hmC is probably less well recognized by Dnmt1 after replication, such that hydroxylation enhances passive demethylation. In contrast to equally distributed hemimethylation we observe a slight increase in the minus strand presence of non-CpG methylation. We cannot find a simple biological (sequence context) or technical (calling/mapping) explanation for this bias. Non-CpG methylation is always occurring in close vicinity to CpG methylation but in contrast to CpGs we find that non-CpGs are not a substrate for Tet enzymes, i.e., we do not find any indication of 5hmC in the non-CpG context. The amount of non-CpG methylation however is strongly enhanced in the absence of Tet enzymes,

which is in line with our observation that *de novo* methylation by Dnmt3a/b is responsible for the non-CpG methylation is enhanced in Tet TKO cells.

Our model provides strong evidence that Dnmts and Tets do not act independently at a given CpG, but clearly in an opposed manner. Generally, we observe a high maintenance and *de novo* efficiency at the majority of the genome, i.e., HMRs and PMDs (or inter-/intragenic regions), while the activity of Tet enzymes is highest at UMRs and LMRs, such as promoters, TFBS (Sox2, Pou5f1) and TSS. Recent studies based on chromatin immunoprecipitation support our findings, revealing binding of Dnmt3a/b at the gene body and HMRs, whereas Tet1 binding was observed across methylation valleys (LMRs and UMRs) [63, 64].

The impairment of maintenance methylation has been identified so far as the main driver of 2i induced DNA demethylation [60] and a role for Tet or oxidative cytosine forms, on the other hand, has only been recognized for selected loci [38, 60]. The comparison of WT and Tet TKO ES cells in the present study, however, disclose a notable reduction within the demethylation rate of Tet TKO, compared to WT ES cells. On average, we detect a reduction in the demethylation rate of almost 50% from around 8% to 4% loss per day. In our view this is enough to demonstrate that Tets and their oxidized cytosine products are essential for an effective demethylation during the Serum-to-2i shift and probably other biological demethylation processes with similar enzymatic compositions.

The loss of Tet enzymes is naturally expected to result in an impaired removal of 5mC and it does at least for CpGs located in LMRs and UMRs, where we observe a notable increase in their methylation level. Nevertheless, under primed conditions and within HMRs we paradoxically observe more unmethylated CpGs (hypomethylation) in Tet TKO ES cells compared to WT ES cells. Recently, López-Moyado *et al.* conducted a systematic investigation of genome wide methylation profiles from various cell types carrying distinct Tet KO genotypes [61]. Similar to our obser-

vations they detected a pronounced loss of DNA methylation in heterochromatic compartments (i.e., HMRs and PMDs) of Tet TKO mouse ES cells.

López-Moyado *et al.* propose a mutual exclusive localization of Dnmts and Tets in WT ES cells, while in Tet KO cells Dnmts invade domains which were previously occupied by Tets. Indeed, in the absence of Tets, our model predicts a clear misregulation in both maintenance and *de novo* methylation efficiency. In Tet TKO ES cells, we see an increase in maintenance methylation efficiency, but at the same time a reduction in HMRs and PMDs. Moreover, we observe an increase in *de novo* methylation efficiency at PMDs. Together, this indicates a displacement of Dnmt1, as well as Dnmt3a/b, which fits to the hypothesized model by López-Moyado *et al.*. In addition, Tet TKO cells exhibit a more stable, almost persistent *de novo* methylation efficiency under naive conditions. The increased non-CpG methylation of Tet TKO cells detected by RRHPoxBS further supports this finding. This shows that in the absence of Tets, ES cells also fail to effectively down-regulate *de novo* methylation efficiency in 2i.

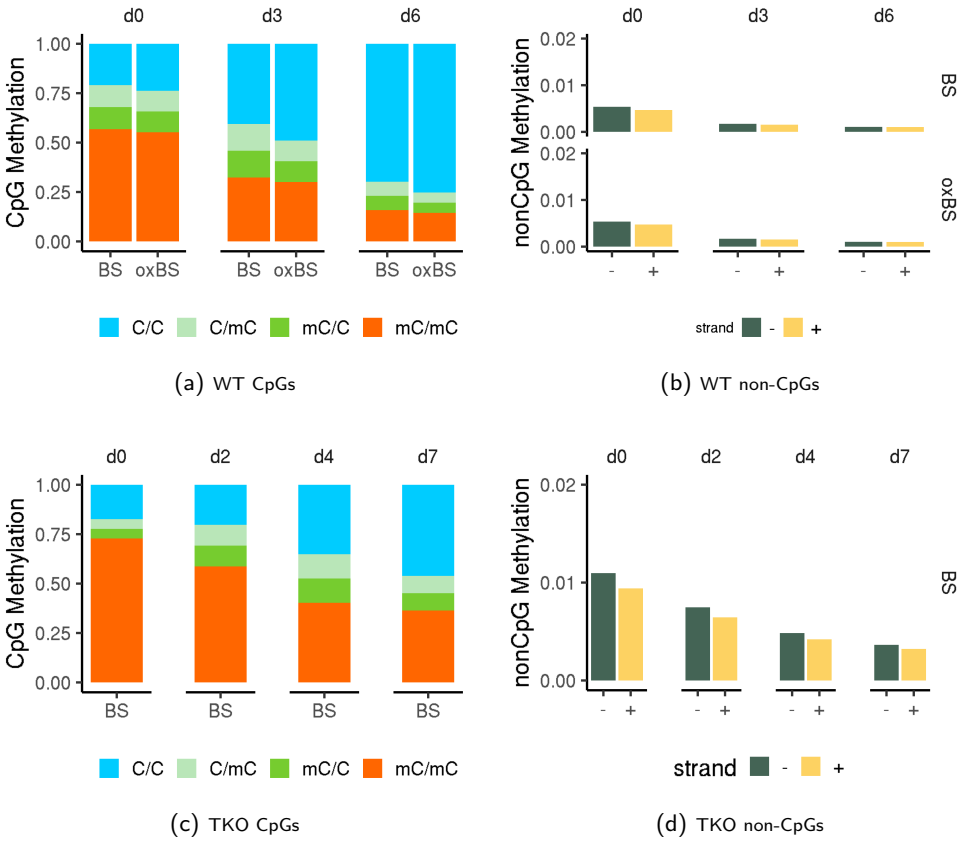
Taken together, we hypothesize that Tet enzymes work against methylation in three ways. (i) They guarantee an efficient conversion of 5mC at accessible regions and act against its establishment during a cell replication either via passive or active demethylation, (ii) They inhibit the effectiveness of the maintenance machinery over regions that should remain unmethylated. (iii) Finally, they ensure an efficient down-regulation of the *de novo* enzymes, which can not be observed in their absence.

Conclusion

We describe a novel combination of experimental and computational approaches to investigate the contributions of Tets and Dnmts to the establishment of distinctive DNA methylation patterns across the genome. Our analysis shows that Dnmts and Tets exhibit clear antagonistic efficiencies at individual CpGs. The comparison

of WT and Tet TKO ES cells demonstrates that Tet enzymes contribute notably to the loss of DNA methylation in the present model system. Moreover, Tet enzymes seem to protect unmethylated regions against both *de novo* and maintenance methylation efficiency and to restrict the activity of Dnmts in highly methylated regions, guaranteeing the formation and maintenance of cell type specific methylation patterns.

Figures



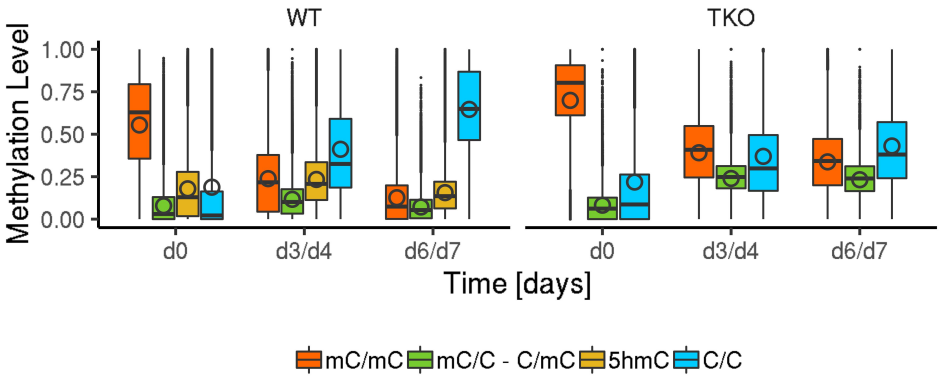
(a) WT CpGs

(b) WT non-CpGs

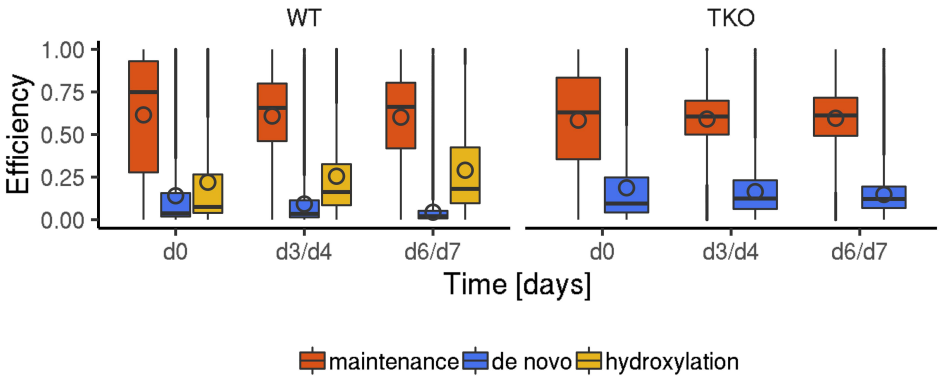
(c) TKO CpGs

(d) TKO non-CpGs

Figure 1: **Demethylation of WT and Tet TKO ES cells. (A)** Average strand specific CpG methylation of WT ES cells in Serum/LIF (d0) and 2i (d3, d6). **(B)** Average strand specific non-CpG methylation of WT ES cells in Serum/LIF and 2i. **(C)** Average strand specific CpG methylation of Tet TKO ES cells in Serum/LIF and 2i (d2, d4, d7). **(D)** Average strand specific non-CpG methylation of Tet TKO ES cells in Serum/LIF and 2i.

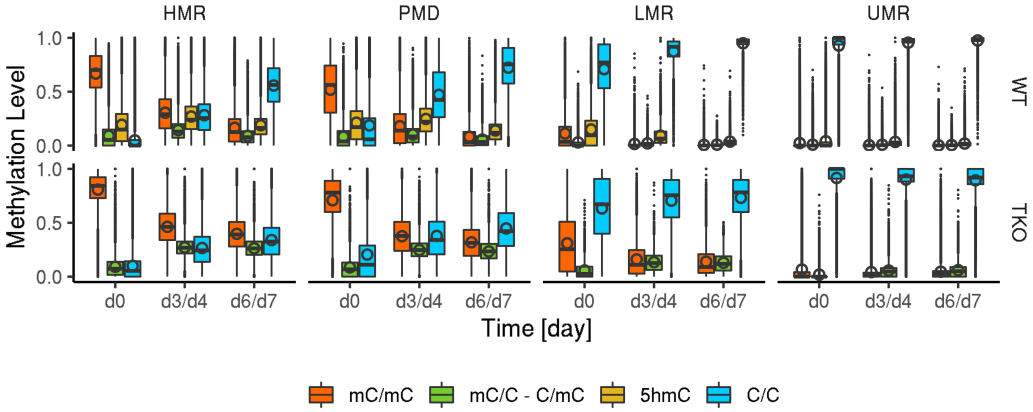


(a) Genome wide estimated methylation levels

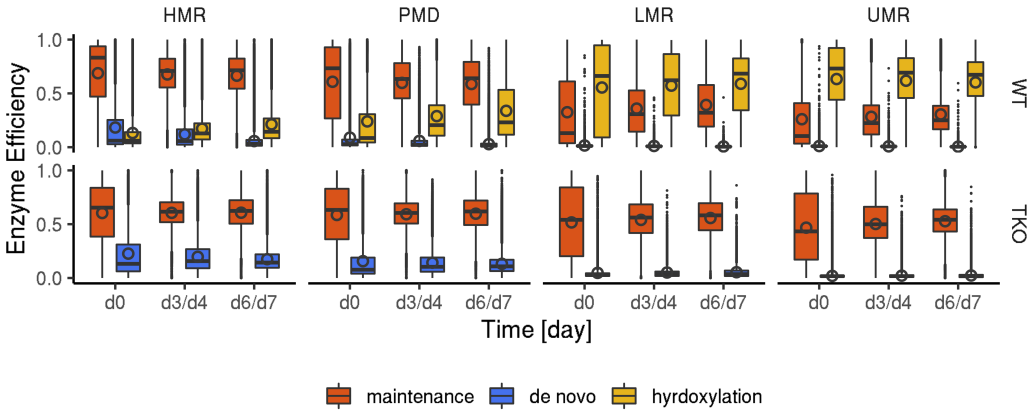


(b) Genome wide estimated enzymatic efficiencies

Figure 2: **Hidden Markov model output. (A)** average estimated 5mC/5hmC level across the genome **(B)** average estimated maintenance efficiency, *de novo* efficiency and hydroxylation efficiency.

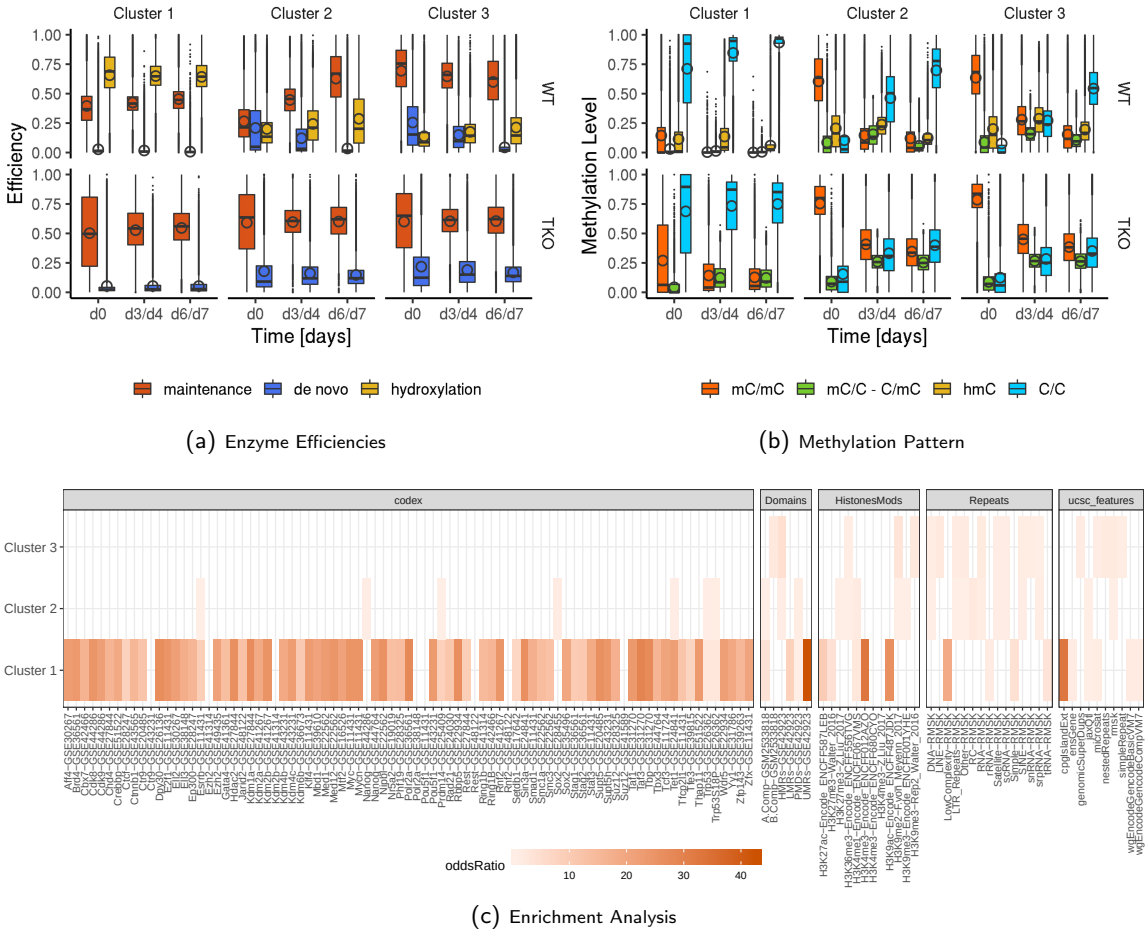


(a) Methylation Pattern



(b) Enzyme Efficiencies

Figure 3: **Methylome Segmentation.** Based on previous published WGBS data, the genome was partitioned in highly methylated regions (HMRs), partially methylated domains (PMDs), low methylated regions (LMRs) and unmethylated regions (UMRs); **(A)** DNA methylation patterns of HMRs, PMDs, LMRs and UMRs derived from HM modeling of RRHPoxBS data; **(B)** HMM estimated enzyme efficiencies of Dnmts (maintenance and *de novo*) and Tets (hydroxylation) for HMRs, PMDs, LMRs and UMRs.



(a) Enzyme Efficiencies

(b) Methylation Pattern

(c) Enrichment Analysis

Figure 4: **Clustering of individual CpGs based on their Efficiency Profile.** CpGs were assigned to three clusters based on their combination of enzyme efficiencies, as well as the changes of enzyme efficiencies over time. **(A)** HMM estimated enzyme efficiencies of Dnmts (maintenance and *de novo*) and Tets (hydroxylation) for Cluster 1 to 3. **(B)** DNA methylation patterns of Cluster 1 to 3 derived from HM modeling of RRHPoxBS data; **(C)** Enrichment analysis of genomic and epigenetic features within the three distinct clusters using LOLA. Colored tiles indicate enrichment with oddsRatio ≥ 1.

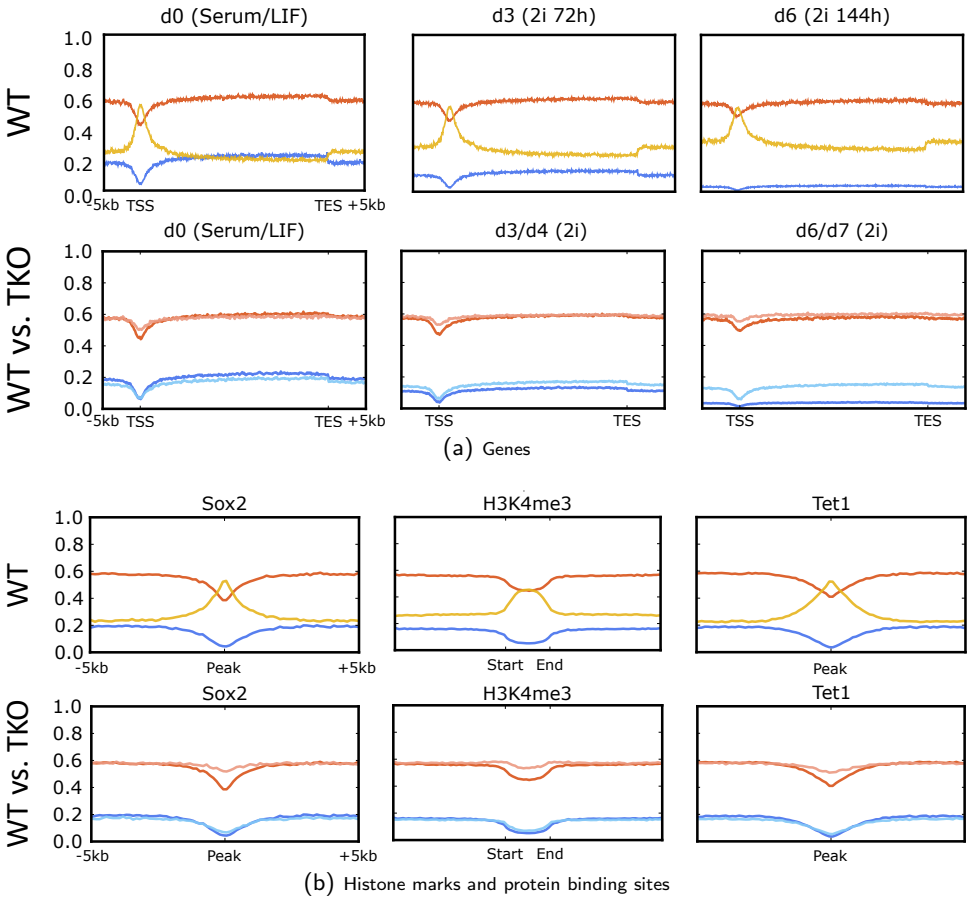


Figure 5: **Average efficiency profiles across genes and protein binding sites.** **(A)** Average maintenance, *de novo* and hydroxylation efficiency of WT and Tet TKO cells across genes; **(B)** Average maintenance, *de novo* and hydroxylation efficiency across selected chromatin marks and protein binding sites. red = maintenance methylation efficiency, blue = *de novo* methylation efficiency, yellow = hydroxylation efficiency. Dark colors indicate the enzyme efficiencies in WT ES cells, light colors refer to the enzyme efficiencies of Tet TKO ES cells.

Competing interests

W.R. is a consultant and shareholder of Cambridge Epigenetix. The remaining authors declare no competing interests.

Funding

The project was funded by the DFG during the course of the SFB1309 and the German Epigenome Programme (DEEP) of the Federal Ministry of Education and Research in Germany (BMBF) [01KU1216 to J.W.].

Author's contributions

Conceived and designed the experiments: CK PG WR VW JW. Wrote the manuscript: CK PG JW VW. Performed the experiments: PG FvM GF. Processed the raw data: KN. Designed/implemented the computational methods: CK. Analyzed the data: CK. Performed meta analysis: CK PG KN AS FM.

Software Availability

All Matlab scripts written for implementing the single CpG stochastic model, run it in a parallel fashion on a multi-core environment and perform the subsequent computational analysis presented in the manuscript and the appendix are shared via GitHub (`https://github.com/kyriakopou/hydroxyGit/tree/master/genomeWide`)

Acknowledgements

We thank Jasmin Kirch and Gilles Gasparoni for performing the sequencing using the Illumina HiSeq2500 system and Andreas Firczynski for the optimization of the

processing pipeline for the hairpin sequencing data.

Supplementary Information

Supplementary_Information.pdf - contains details about modeling, the parameter estimation method, as well as further computational analysis and additional results.

SI_Table1.pdf - contains the complete output of the LOLA enrichment analysis.

Accession Numbers

The data will be available at the NCBI GEO database

References

[1] Robin Holliday and John E Pugh. Dna modification mechanisms and gene activity during development. *Science*, 187(4173):226–232, 1975.

[2] Arthur D Riggs. X inactivation, differentiation, and dna methylation. *Cytogenetic and Genome Research*, 14(1):9–25, 1975.

[3] Déborah Bourc'his and Timothy H Bestor. Meiotic catastrophe and retrotransposon reactivation in male germ cells lacking dnmt3l. *Nature*, 431(7004):96, 2004.

[4] En Li, Timothy H Bestor, and Rudolf Jaenisch. Targeted mutation of the dna methyltransferase gene results in embryonic lethality. *Cell*, 69(6):915–926, 1992.

[5] Bernard H Ramsahoye, Detlev Biniszkiewicz, Frank Lyko, Victoria Clark, Adrian P Bird, and Rudolf Jaenisch. Non-cpg methylation is prevalent in embryonic stem cells and may be mediated by dna methyltransferase 3a. *Proceedings of the National Academy of Sciences*, 97(10):5237–5242, 2000.

[6] Michael J Ziller, Fabian Müller, Jing Liao, Yingying Zhang, Hongcang Gu, Christoph Bock, Patrick Boyle, Charles B Epstein, Bradley E Bernstein, Thomas Lengauer, et al. Genomic distribution and inter-sample variation of non-cpg methylation across human cell types. *PLoS genetics*, 7(12):e1002389, 2011.

[7] Ryan Lister, Mattia Pelizzola, Robert H Dowen, R David Hawkins, Gary Hon, Julian Tonti-Filippini, Joseph R Nery, Leonard Lee, Zhen Ye, Que-Minh Ngo, et al. Human dna methylomes at base resolution show widespread epigenomic differences. *nature*, 462(7271):315, 2009.

[8] Heinrich Leonhardt, Andrea W Page, Heinz-Ulrich Weier, and Timothy H Bestor. A targeting sequence directs dna methyltransferase to sites of dna replication in mammalian nuclei. *Cell*, 71(5):865–873, 1992.

[9] Linda S-H Chuang, Hang-In Ian, Tong-Wey Koh, Huck-Hui Ng, Guoliang Xu, and Benjamin FL Li. Human dna-(cytosine-5) methyltransferase-pcna complex as a target for p21waf1. *Science*, 277(5334):1996–2000, 1997.

[10] Magnolia Bostick, Jong Kyong Kim, Pierre-Olivier Estève, Amander Clark, Sriharsa Pradhan, and Steven E Jacobsen. Uhrf1 plays a role in maintaining dna methylation in mammalian cells. *Science*, 317(5845):1760–1764, 2007.

[11] Jafar Sharif, Masahiro Muto, Shin-ichiro Takebayashi, Isao Suetake, Akihiro Iwamatsu, Takaho A Endo, Jun Shinga, Yoko Mizutani-Koseki, Tetsuro Toyoda, Kunihiro Okamura, et al. The sra protein np95 mediates epigenetic inheritance by recruiting dnmt1 to methylated dna. *Nature*, 450(7171):908–912, 2007.

[12] Kyohei Arita, Mariko Ariyoshi, Hidehito Tochio, Yusuke Nakamura, and Masahiro Shirakawa. Recognition of hemi-methylated dna by the sra protein uhrf1 by a base-flipping mechanism. *Nature*, 455(7214):818, 2008.

[13] Andrea Hermann, Rachna Goyal, and Albert Jeltsch. The dnmt1 dna-(cytosine-c5)-methyltransferase methylates dna processively with high preference for hemimethylated target sites. *Journal of Biological Chemistry*, 279(46):48350–48359, 2004.

[14] Masaki Okano, Shaoping Xie, and En Li. Cloning and characterization of a family of novel mammalian dna (cytosine-5) methyltransferases. *Nature genetics*, 19(3):219, 1998.

[15] Masaki Okano, Daphne W Bell, Daniel A Haber, and En Li. Dna methyltransferases dnmt3a and dnmt3b are essential for de novo methylation and mammalian development. *Cell*, 99(3):247–257, 1999.

[16] Daniela Meilinger, Karin Fellinger, Sebastian Bultmann, Ulrich Rothbauer, Ian Marc Bonapace, Wolfgang EF Klinkert, Fabio Spada, and Heinrich Leonhardt. Np95 interacts with de novo dna methyltransferases, dnmt3a and dnmt3b, and mediates epigenetic silencing of the viral cmv promoter in embryonic stem cells. *EMBO reports*, 10(11):1259–1264, 2009.

[17] Gangning Liang, Matilda F Chan, Yoshitaka Tomigahara, Yvonne C Tsai, Felicidad A Gonzales, En Li, Peter W Laird, and Peter A Jones. Cooperativity between dna methyltransferases in the maintenance methylation of repetitive elements. *Molecular and cellular biology*, 22(2):480–491, 2002.

[18] Julia Arand, David Spieler, Tommy Karius, Miguel R Branco, Daniela Meilinger, Alexander Meissner, Thomas Jenuwein, Guoliang Xu, Heinrich Leonhardt, Verena Wolf, et al. In vivo control of cpg and non-cpg dna methylation by dna methyltransferases. *PLoS genetics*, 8(6):e1002750, 2012.

[19] Ryoichi Ono, Tomohiko Taki, Takeshi Taketani, Masafumi Taniwaki, Hajime Kobayashi, and Yasuhide Hayashi. Lcx, leukemia-associated protein with a cxxc domain, is fused to mll in acute myeloid leukemia with trilineage dysplasia having t (10; 11)(q22; q23). *Cancer research*, 62(14):4075–4080, 2002.

[20] RB Lorsbach, J Moore, S Mathew, SC Raimondi, ST Mukatira, and JR Downing. Tet1, a member of a novel protein family, is fused to mll in acute myeloid leukemia containing the t (10; 11)(q22; q23). *Leukemia*, 17(3):637, 2003.

[21] Lakshminarayan M Iyer, Mamta Tahiliani, Anjana Rao, and L Aravind. Prediction of novel families of enzymes involved in oxidative and other complex modifications of bases in nucleic acids. *Cell cycle*, 8(11):1698–1710, 2009.

[22] Mamta Tahiliani, Kian Peng Koh, Yinghua Shen, William A Pastor, Hozefa Bandukwala, Yevgeny Brudno, Suneet Agarwal, Lakshminarayan M Iyer, David R Liu, L Aravind, et al. Conversion of 5-methylcytosine to 5-hydroxymethylcytosine in mammalian dna by mll partner tet1. *Science*, 324(5929):930–935, 2009.

[23] Shinsuke Ito, Li Shen, Qing Dai, Susan C Wu, Leonard B Collins, James A Swenberg, Chuan He, and Yi Zhang. Tet proteins can convert 5-methylcytosine to 5-formylcytosine and 5-carboxylcytosine. *Science*, 333(6047):1300–1303, 2011.

[24] Daniel Globisch, Martin Münzel, Markus Müller, Stylianos Michalakis, Mirko Wagner, Susanne Koch, Tobias Brückl, Martin Biel, and Thomas Carell. Tissue distribution of 5-hydroxymethylcytosine and search for active demethylation intermediates. *PloS one*, 5(12):e15367, 2010.

[25] Skirmantas Kriaucionis and Nathaniel Heintz. The nuclear dna base 5-hydroxymethylcytosine is present in purkinje neurons and the brain. *Science*, 324(5929):929–930, 2009.

[26] Aleksandra Szwagierczak, Sebastian Bultmann, Christine S Schmidt, Fabio Spada, and Heinrich Leonhardt. Sensitive enzymatic quantification of 5-hydroxymethylcytosine in genomic dna. *Nucleic acids research*, 38(19):e181–e181, 2010.

[27] Martin Bachman, Santiago Uribe-Lewis, Xiaoping Yang, Michael Williams, Adele Murrell, and Shankar Balasubramanian. 5-hydroxymethylcytosine is a predominantly stable dna modification. *Nature chemistry*, 6(12):1049, 2014.

[28] Eun-Ang Raiber, Pierre Murat, Dimitri Y Chirgadze, Dario Beraldi, Ben F Luisi, and Shankar Balasubramanian. 5-formylcytosine alters the structure of the dna double helix. *Nature Structural and Molecular Biology*, 22(1):44, 2015.

[29] Matthew W Kellinger, Chun-Xiao Song, Jenny Chong, Xing-Yu Lu, Chuan He, and Dong Wang. 5-formylcytosine and 5-carboxylcytosine reduce the rate and substrate specificity of rna polymerase ii transcription. *Nature Structural and Molecular Biology*, 19(8):831, 2012.

[30] Hideharu Hashimoto, Yiwei Liu, Anup K Upadhyay, Yanqi Chang, Shelley B Howerton, Paula M Vertino, Xing Zhang, and Xiaodong Cheng. Recognition and potential mechanisms for replication and erasure of cytosine hydroxymethylation. *Nucleic acids research*, 40(11):4841–4849, 2012.

[31] Victoria Valinluck and Lawrence C Sowers. Endogenous cytosine damage products alter the site selectivity of human dna maintenance methyltransferase dnmt1. *Cancer research*, 67(3):946–950, 2007.

[32] Debin Ji, Krystal Lin, Jikui Song, and Yinsheng Wang. Effects of tet-induced oxidation products of 5-methylcytosine on dnmt1-and dnmt3a-mediated cytosine methylation. *Molecular bioSystems*, 10(7):1749–1752, 2014.

[33] Yu-Fei He, Bin-Zhong Li, Zheng Li, Peng Liu, Yang Wang, Qingyu Tang, Jianping Ding, Yingying Jia, Zhangcheng Chen, Lin Li, et al. Tet-mediated formation of 5-carboxylcytosine and its excision by tdg in mammalian dna. *Science*, 333(6047):1303–1307, 2011.

[34] Atanu Maiti and Alexander C Drohat. Thymine dna glycosylase can rapidly excise 5-formylcytosine and 5-carboxylcytosine potential implications for active demethylation of cpg sites. *Journal of Biological Chemistry*, 286(41):35334–35338, 2011.

[35] Zachary D Smith, Michelle M Chan, Tarjei S Mikkelsen, Hongcang Gu, Andreas Gnirke, Aviv Regev, and Alexander Meissner. A unique regulatory phase of dna methylation in the early mammalian embryo. *Nature*, 484(7394):339, 2012.

[36] J Oswald, S Engemann, N Lane, W Mayer, A Olek, R Fundele, W Dean, W Reik, and J Walter. Active demethylation of the paternal genome in the mouse zygote. *Current Biology*, 10(8):475–478, 2000.

[37] Petra Hajkova, Sean J Jeffries, Caroline Lee, Nigel Miller, Stephen P Jackson, and M Azim Surani. Genome-wide reprogramming in the mouse germ line entails the base excision repair pathway. *Science*, 329(5987):78–82, 2010.

[38] Gabriella Ficz, Timothy A Hore, Fatima Santos, Heather J Lee, Wendy Dean, Julia Arand, Felix Krueger, David Oxley, Yu-Lee Paul, Jörn Walter, et al. Fgf signaling inhibition in escs drives rapid genome-wide demethylation to the epigenetic ground state of pluripotency. *Cell stem cell*, 13(3):351–359, 2013.

[39] Ehsan Habibi, Arie B Brinkman, Julia Arand, Leonie I Kroeze, Hindrik HD Kerstens, Filomena Matarese, Konstantin Lepikhov, Marta Gut, Isabelle Brun-Heath, Nina C Hubner, et al. Whole-genome bisulfite sequencing of two distinct interconvertible dna methylomes of mouse embryonic stem cells. *Cell stem cell*, 13(3):360–369, 2013.

[40] Marius Walter, Aurélie Teissandier, Raquel Pérez-Palacios, and Déborah Bourc'his. An epigenetic switch ensures transposon repression upon dynamic loss of dna methylation in embryonic stem cells. *Elife*, 5, 2016.

[41] Alexander Meissner, Andreas Gnirke, George W Bell, Bernard Ramsahoye, Eric S Lander, and Rudolf Jaenisch. Reduced representation bisulfite sequencing for comparative high-resolution dna methylation analysis. *Nucleic acids research*, 33(18):5868–5877, 2005.

[42] Charles D Laird, Nicole D Pleasant, Aaron D Clark, Jessica L Sneeden, KM Anwarul Hassan, Nathan C Manley, Jay C Vary, Todd Morgan, R Scott Hansen, and Reinhard Stöger. Hairpin-bisulfite pcr: assessing epigenetic methylation

patterns on complementary strands of individual dna molecules. *Proceedings of the National Academy of Sciences*, 101(1):204–209, 2004.

[43] Pascal Giehr and Jörn Walter. Hairpin bisulfite sequencing: Synchronous methylation analysis on complementary dna strands of individual chromosomes. In *DNA Methylation Protocols*, pages 573–586. Springer, 2018.

[44] Lei Zhao, Ming-an Sun, Zejuan Li, Xue Bai, Miao Yu, Min Wang, Liji Liang, Xiaojian Shao, Stephen Arnovitz, Qianfei Wang, et al. The dynamics of dna methylation fidelity during mouse embryonic stem cell self-renewal and differentiation. *Genome research*, 24(8):1296–1307, 2014.

[45] Michael J Booth, Miguel R Branco, Gabriella Ficz, David Oxley, Felix Krueger, Wolf Reik, and Shankar Balasubramanian. Quantitative sequencing of 5-methylcytosine and 5-hydroxymethylcytosine at single-base resolution. *Science*, 336(6083):934–937, 2012.

[46] Charalampos Kyriakopoulos, Pascal Giehr, and Verena Wolf. H (o) ta: estimation of dna methylation and hydroxylation levels and efficiencies from time course data. *Bioinformatics*, 33(11):1733–1734, 2017.

[47] Pascal Giehr, Charalampos Kyriakopoulos, Gabriella Ficz, Verena Wolf, and Jörn Walter. The influence of hydroxylation on maintaining cpg methylation patterns: a hidden markov model approach. *PLoS computational biology*, 12(5):e1004905, 2016.

[48] Jacob Porter, Ming-an Sun, Hehuang Xie, and Liqing Zhang. Investigating bisulfite short-read mapping failure with hairpin bisulfite sequencing data. *BMC genomics*, 16(11):S2, 2015.

[49] Babraham Bioinformatics - Trim Galore!

[50] Marcel Martin. Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet. journal*, 17(1):pp–10, 2011.

[51] Santiago Marco-Sola, Michael Sammeth, Roderic Guigó, and Paolo Ribeca. The gem mapper: fast, accurate and versatile alignment by filtration. *Nature methods*, 9(12):1185, 2012.

[52] Heng Li, Bob Handsaker, Alec Wysoker, Tim Fennell, Jue Ruan, Nils Homer, Gabor Marth, Goncalo Abecasis, and Richard Durbin. The sequence alignment/map format and samtools. *Bioinformatics*, 25(16):2078–2079, 2009.

[53] Leo Goodstadt. Ruffus: a lightweight python library for computational pipelines. *Bioinformatics*, 26(21):2778–2779, 2010.

[54] Samuel L Braunstein. How large a sample is needed for the maximum likelihood estimator to be approximately gaussian? *Journal of Physics A: Mathematical and General*, 25(13):3813, 1992.

[55] Scott J Long and Jeremy Freese. *Regression models for categorical dependent variables using Stata*. Stata press, 2006.

[56] Ronald Schoenberg. Constrained maximum likelihood. *Computational Economics*, 10(3):251–266, 1997.

[57] Lukas Burger, Dimos Gaidatzis, Dirk Schübeler, and Michael B Stadler. Identification of active regulatory regions from dna methylation data. *Nucleic acids research*, 41(16):e155–e155, 2013.

[58] Cath Tyner, Galt P Barber, Jonathan Casper, Hiram Clawson, Mark Diekhans, Christopher Eisenhart, Clayton M Fischer, David Gibson, Jairo Navarro Gonzalez, Luvina Guruvadoo, et al. The ucsc genome browser database: 2017 update. *Nucleic acids research*, 45(D1):D626–D634, 2016.

[59] Nathan C Sheffield and Christoph Bock. Lola: enrichment analysis for genomic region sets and regulatory elements in r and bioconductor. *Bioinformatics*, 32(4):587–589, 2015.

[60] Ferdinand von Meyenn, Mario Iurlaro, Ehsan Habibi, Ning Qing Liu, Ali Salehzadeh-Yazdi, Fátima Santos, Edoardo Petrini, Inês Milagre, Miao Yu, Zhenqing Xie, et al. Impairment of dna methylation maintenance is the main cause of global demethylation in naive embryonic stem cells. *Molecular cell*, 62(6):848–861, 2016.

[61] Isaac F López-Moyado, Ageliki Tsagaratou, Hiroshi Yuita, Hyungseok Seo, Benjamin Delatte, Sven Heinz, Christopher Benner, and Anjana Rao. Paradoxical association of tet loss of function with genome-wide dna hypomethylation. *Proceedings of the National Academy of Sciences*, 116(34):16933–16942, 2019.

[62] Fidel Ramírez, Friederike Dündar, Sarah Diehl, Björn A Grüning, and Thomas Manke. deeptools: a flexible platform for exploring deep-sequencing data. *Nucleic acids research*, 42(W1):W187–W191, 2014.

[63] Tuncay Baubec, Daniele F Colombo, Christiane Wirbelauer, Juliane Schmidt, Lukas Burger, Arnaud R Krebs, Altuna Akalin, and Dirk Schübeler. Genomic profiling of dna methyltransferases reveals a role for dnmt3b in genic methylation. *Nature*, 520(7546):243, 2015.

[64] Tianpeng Gu, Xueqiu Lin, Sean M Cullen, Min Luo, Mira Jeong, Marcos Estecio, Jianjun Shen, Swanand Hardikar, Deqiang Sun, Jianzhong Su, et al. Dnmt3a and tet1 cooperate to regulate promoter epigenetic landscapes in mouse embryonic stem cells. *Genome biology*, 19(1):88, 2018.