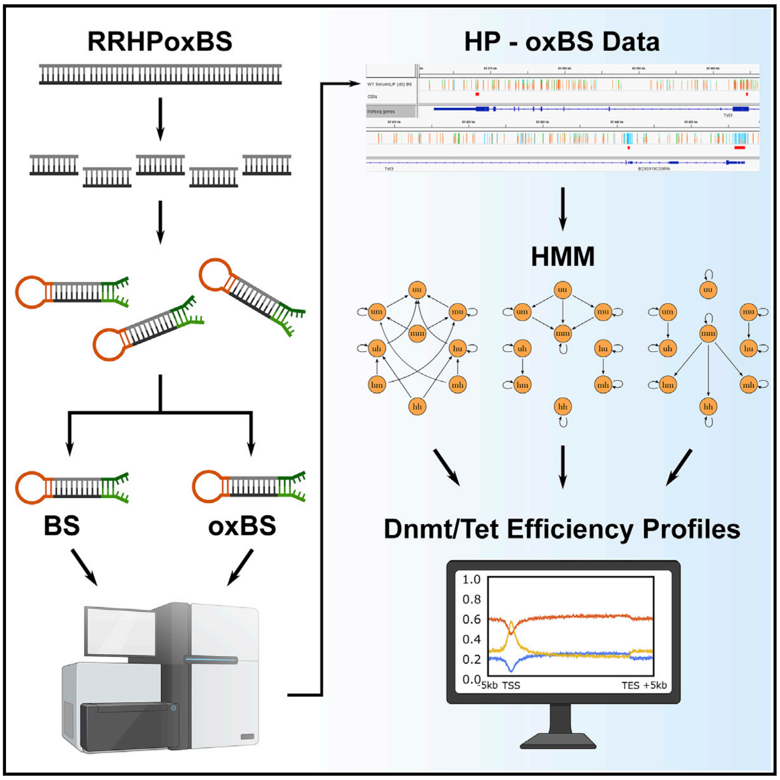
Cell Reports
Methods

Article

A comprehensive approach for genome-wide efficiency profiling of DNA modifying enzymes

Graphical abstract



Authors

Charalampos Kyriakopoulos,
Karl Nordström, Paula Linh Kramer, ...,
Verena Wolf, Jörn Walter, Pascal Giehr

Correspondence

pgiehr@ethz.ch

In brief

Kyriakopoulos et al. develop a pipeline for quantitative estimation of Dnmt and Tet activity. Using double-strand methylation information, GwEEP infers maintenance and *de novo* methylation efficiency of Dnmts as well as hydroxylation efficiency of Tets at single-base resolution.

Highlights

- Genome-wide, strand-specific oxidative hairpin sequencing
- Precise temporal estimation of 5mC/5hmC distribution and Dnmt/Tet activity
- Tets presence contributes notably to DNA demethylation
- Model output validates the mutual interference of Dnmts and Tets

Kyriakopoulos et al., 2022, Cell Reports Methods *2*, 100187
March 28, 2022 © 2022 The Authors.
https://doi.org/10.1016/j.crmeth.2022.100187

CellPress

Cell Reports Methods

CellPress
OPEN ACCESS

Article

A comprehensive approach for genome-wide efficiency profiling of DNA modifying enzymes

Charalampos Kyriakopoulos,[1,8] Karl Nordström,[2,9] Paula Linh Kramer,[1] Judith Yumiko Gottfreund,[2] Abdulrahman Salhab,[2] Julia Arand,[3] Fabian Müller,[4] Ferdinand von Meyenn,[5] Gabriella Ficz,[6] Wolf Reik,[7] Verena Wolf,[1] Jörn Walter,[2] and Pascal Giehr[2,5,10,*]

[1] Computer Science Department, Saarland University, Campus E1.3, 66123 Saarbrücken, Germany
[2] Department of Genetics and Epigenetics, Saarland University, Campus A2.4, 66123 Saarbrücken, Germany
[3] Division of Cell and Developmental Biology, Medical University of Vienna, 1090 Vienna, Austria
[4] Department of Integrative Cellular Biology and Bioinformatics, Campus A2.4, 66123 Saarbrücken, Germany
[5] Department of Health Sciences and Technology, ETH Zürich, Schorenstrasse 16, Schwerzenbach, 8603 Zürich, Switzerland
[6] Haemato-Oncology, Queen Mary University of London, London EC1M 6BQ, UK
[7] Epigenetics Department, Babraham Institute, Cambridge CB22 3AT, UK
[8] Present address: Bristol Myers Squibb, Center for Innovation and Translational Research Europe, Calle Isaac Newton 4, 41092 Sevilla, Spain
[9] Present address: Astra Zeneca, Pepparedsleden 1, 431 50 Mölndal, Sweden
[10] Lead contact
*Correspondence: pgiehr@ethz.ch

https://doi.org/10.1016/j.crmeth.2022.100187

MOTIVATION Dynamic changes of DNA methylation patterns are a common phenomenon in epigenetics. Although a stable DNA methylation profile is essential for cell identity, developmental processes require the rearrangement of 5-methylcytosine in the genome. Stable methylation patterns are the result of balanced Dnmts and Tets activities, while methylome transformation results from a coordinated change in Dnmt and Tet efficiencies. Such transformations occur on a global scale, e.g., during the reprogramming of maternal and paternal methylation patterns and the establishment of novel cell-type-specific methylomes during embryonic development *in vivo*, but also *in vitro* during (re)programming of induced pluripotent stem cells as well as somatic cells. In addition, local (de)methylation events are the key to gene regulation during cell differentiation. A detailed characterization of Dnmt and Tet cooperation is essential for understanding natural epigenetic adaptation as well as the optimization of *in vitro* (re)programming protocols. For this purpose we developed a pipeline for quantitative and precise estimation of Dnmt and Tet activity. Using only double-strand methylation information GwEEP infers accurate maintenance and *de novo* methylation efficiency of Dnmts as well as hydroxylation efficiency of Tets at single-base resolution. Thus, we believe GwEEP provides a powerful tool for the investigation of methylome rearrangements in various systems.

SUMMARY

A precise understanding of DNA methylation dynamics is of great importance for a variety of biological processes including cellular reprogramming and differentiation. To date, complex integration of multiple and distinct genome-wide datasets is required to realize this task. We present GwEEP (genome-wide epigenetic efficiency profiling) a versatile approach to infer dynamic efficiencies of DNA modifying enzymes. GwEEP relies on genome-wide hairpin datasets, which are translated by a hidden Markov model into quantitative enzyme efficiencies with reported confidence around the estimates. GwEEP predicts *de novo* and maintenance methylation efficiencies of Dnmts and furthermore the hydroxylation efficiency of Tets. Its design also allows capturing further oxidation processes given available data. We show that GwEEP predicts accurately the epigenetic changes of ESCs following a Serum-to-2i shift and applied to Tet TKO cells confirms the hypothesized mutual interference between Dnmts and Tets.

INTRODUCTION

Genetic information encoded in the DNA is regulated by epigenetic mechanisms, such as DNA methylation (Holliday and Pugh, 1975; Riggs, 1975; Bourc'his and Bestor, 2004; Li et al., 1992). In mammals methylation of DNA is restricted to cytosine and it is almost exclusively found in a palindromic CpG dinucleotide context (Ramsahoye et al., 2000; Ziller et al., 2011; Lister et al., 2009). Generation of 5-methylcytosine (5mC) is catalyzed by the DNA methyltransferases Dnmt1, Dnmt3a and Dnmt3b. These enzymes catalyze the transfer of a methyl group from S-adenosyl methionine to the fifth carbon atom of cytosine. Dnmt1 is responsible for maintaining existing methylation patterns after replication. Via interaction with Uhrf1 and PCNA, Dnmt1 is tightly associated with the replication machinery (Leonhardt et al., 1992; Chuang et al., 1997). Furthermore, the cooperation with Uhrf1 modulates Dnmt1 to be receptive for hemimethylated DNA generated after replication (Bostick et al., 2007; Sharif et al., 2007) and thus the protein complex post-replicatively copies the methylation pattern from the inherited to the newly synthesized DNA strand (Arita et al., 2008; Hermann et al., 2004). Dnmt3a and Dnmt3b methylate DNA independently of its methylation status (hemimethylated or unmethylated) and are mainly responsible for the establishment of new methylation patterns during development (Okano et al., 1998, 1999). However, several studies indicate that the strict separation of Dnmt1 and Dnmt3a/b activity is not coherent and that under certain conditions these enzymes exhibit overlapping functions (Meilinger et al., 2009; Liang et al., 2002; Arand et al., 2012).

Once established, 5mC can be further processed by a family of di-oxigenases, the ten-eleven translocation enzymes Tet1, Tet2, and Tet3 (Ono et al., 2002; Lorsbach et al., 2003; Iyer et al., 2009). These Fe(II)- and oxoglutarate-dependent enzymes consecutively oxidize 5mC to 5-hydroxymethyl cytosine (5hmC), 5-formyl cytosine (5fC) and ultimately to 5-carboxy cytosine (5caC) (Tahiliani et al., 2009; Ito et al., 2011). Each oxidation step changes the chemical properties of the base and with it its biological function (Bachman et al., 2014; Raiber et al., 2015; Kellinger et al., 2012), albeit 5hmC is definitely the most abundant oxidative variant found in numerous cell types (Globisch et al., 2010; Kriaucionis and Heintz, 2009; Szwagierczak et al., 2010). Several mechanisms have been proposed in which oxidative cytosine derivatives (oxC) serve as an intermediate during the course of active or passive demethylation (Hashimoto et al., 2012; Valinluck and Sowers, 2007; Ji et al., 2014; He et al., 2011; Maiti and Drohat, 2011). Such removal of 5mC occurs locally during cell differentiation, but also on a genome-wide scale in the zygote or during the maturation of primordial germ cells (Smith et al., 2012; Oswald et al., 2000; Hajkova et al., 2010). Genome-wide loss of 5mC has also been observed in cultivated mouse embryonic stem cells (ESCs) during their transition from Serum to 2i medium. Under classical serum/LIF conditions ESCs exhibit DNA hypermethylation, whereas upon transition to GSK3 and Erk1/2 inhibitors (2i) containing medium the cells experience a gradual genome-wide loss of 5mC (Ficz et al., 2013; Habibi et al., 2013; Walter et al., 2016).

Even though several studies have examined the influence of Tets and oxCs (Gu et al., 2018; López-Moyado et al., 2019; Ginno et al., 2020; Charlton et al., 2020) within the genome, the precise contribution of Tets and oxCs toward maintaining or changing cell-type-specific methylomes remains elusive. However, a thorough understanding of the local and spatial connections between Dnmts and Tets during the processes of development, cell division and differentiation is of great importance as it can form the basis for a structured development of novel epigenetic cancer therapies and/or controlled reprogramming approaches in regenerative stem cell medicine. To contribute to addressing this complex interplay between Dnmts and Tets we developed the following experimental and computational pipeline.

RESULTS

GwEEP (genome-wide epigenetic efficiency profiling) consists of three major parts: (1) The construction of a hairpin oxidative bisulfite library - an endonuclease-based enrichment of representative CpGs (Meissner et al., 2005), which we named reduced representation hairpin oxidative bisulfite sequencing (RRHPoxBS), (2) a computational pipeline (HPup) which extracts the double-strand DNA methylation values from Illumina sequencing data, and (3) a hidden Markov model that estimates conversion error corrected 5mC and 5hmC distributions and furthermore infers the enzymatic efficiencies of Dnmts (maintenance and *de novo* methylation efficiency) as well as Tets (hydroxylation efficiency). The pipeline is outlined in Figure 1 and the details of the individual steps are described in the STAR Methods.

We applied GwEEP on a well-established ESC system to precisely map 5mC and 5hmC across the genome in a time series experiment and studied the enzymatic contribution of Tets and Dnmts for the progressive genome-wide DNA (de)methylation. To achieve this we first generated a high-resolution dataset based on the above-described genome-wide hairpin sequencing approach, RRHPoxBS. The design of the RRHPoxBS (combination of R.AluI, R.HaeIII, and R.HpyCH4V) approach covers around 4 million CpG dyads (16.3%–22.5% of the genome; Table S1) located in CpG-poor and -rich regions 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 datasets for WT ESCs, i.e., bisulfite (BS) and oxidative bisulfite (oxBS) libraries for three different time points, starting with serum/Lif (d0), followed by 72 h 2i (d3) and 144 h 2i (d6). For a comparison we also generated four datasets for Tet TKO cells starting with serum/Lif (d0) followed by 48 h in 2i (d2), 96 h in 2i (d4), and 168 h in 2i (d7).

Impaired loss of 5mC in Tet TKO ESCs

In contrast to canonical bisulfite sequencing (Ficz et al., 2013; Habibi et al., 2013; von Meyenn et al., 2016), RRHPoxBS allow us to additionally identify hemimethylated CpGs and therefore precisely estimate the demethylation kinetics $\left(r_{\text{dem}}(t) = \frac{\text{TT}(t)-\text{TT}(0)}{t}\right)$ at day *t* revealing that in WT ESCs the generation of unmethylated cytosine is 8% per day, while in Tet TKO cells it drops to 4.2% per day (Figures 2C and 2F).

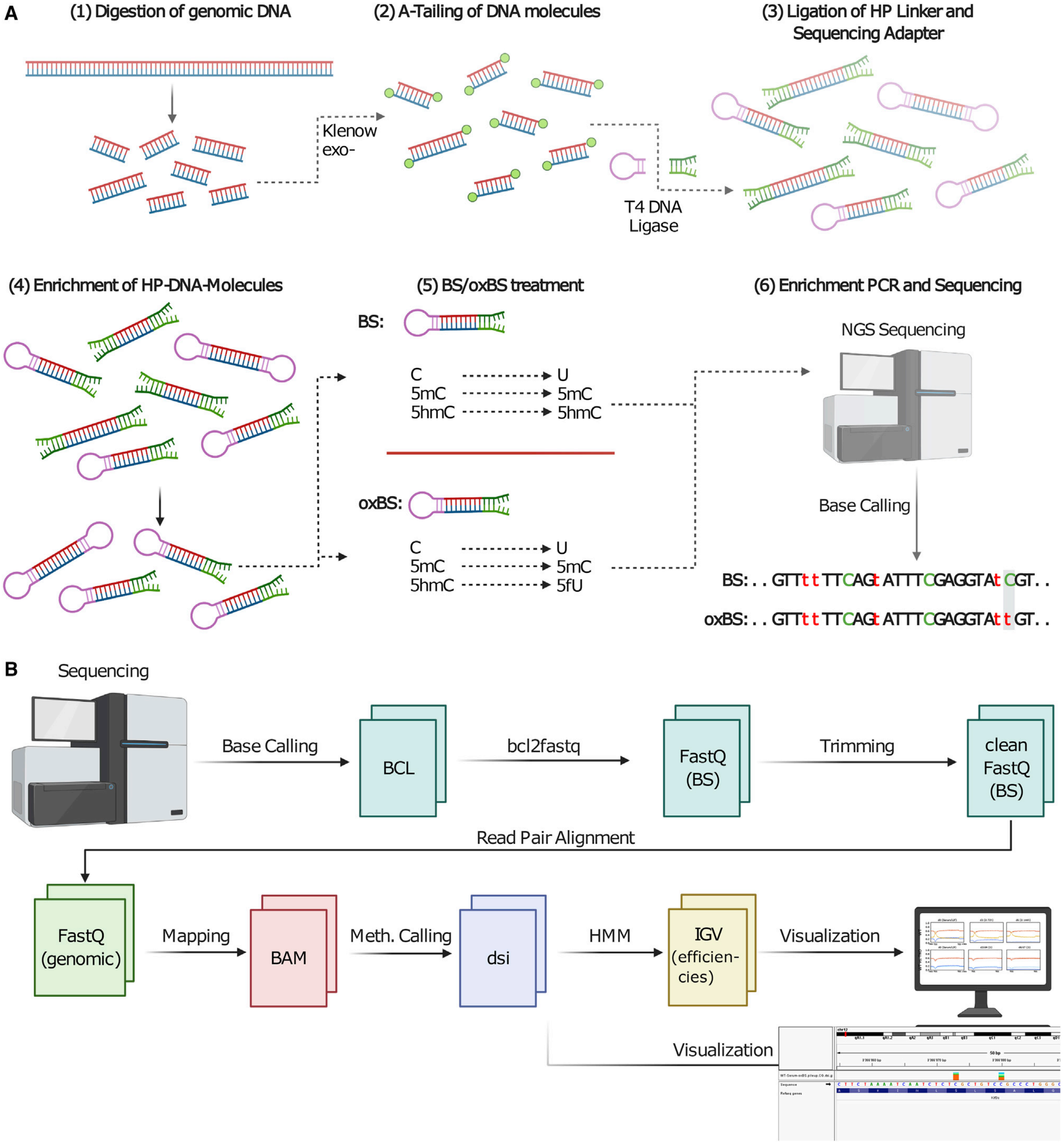


Figure 1. GwEEP - Pipeline overview

(A) Laboratory pipeline: (1) Genomic DNA is digested by endo nucleases followed by (2) Klenow exo-catalyzed A-tailing. (3) A-Tailed DNA molecules are subjected to sequencing adapter, hairpin linker ligation and (4) subsequent enrichment of hairpin-ligated molecules. (5) Half of the library is used for BS, the other half for oxBS treatment. (6) After amplification and indexing using PCR the libraries are sequenced on an Illumina platform with minimum 100 bp in a paired-end mode (created with BioRender.com).

(B) Computational processing: Illumina raw data are processed into base calls (FASTQ) and trimmed for adapter and hairpin linker sequences. Bisulfite reads from the same molecules are paired to restore the genomic sequence for efficient mapping. Subsequently, the double-strand information is annotated and stored in DSI (double strand information) files. The HMM then derives 5mC and 5hmC distributions, as well as the efficiencies of Dnmt and Tets which are stored in the IGV file format. Both DSI and IGV files can be visualized using the IGV genome browser. (created with BioRender.com).

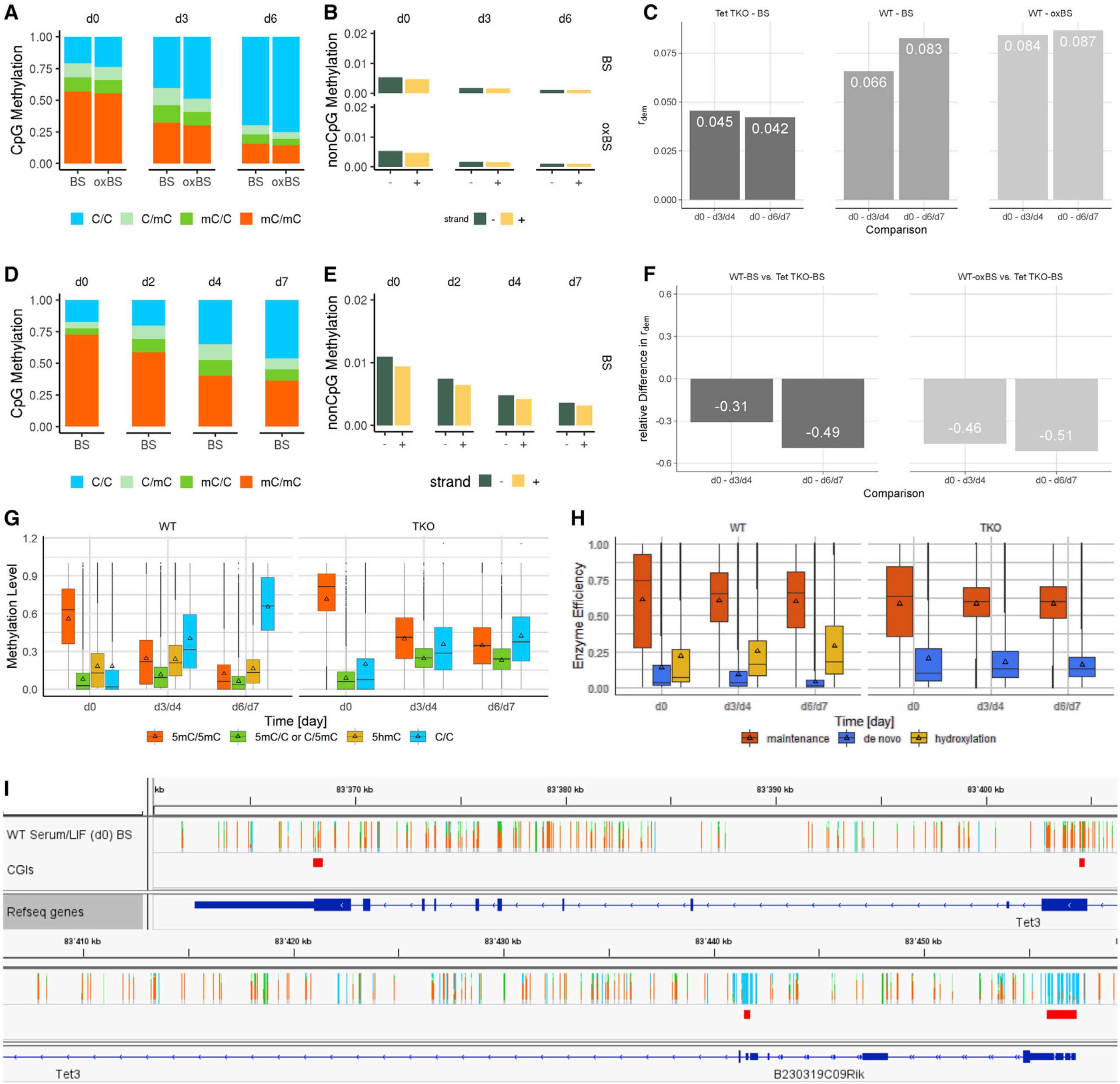


Figure 2. RRoxBS and HMM results

(A) Average CpG methylation level based on uncorrected hairpin sequencing counts for WT ESCs.
(B) Average nonCpG methylation level based on uncorrected hairpin sequencing counts for WT ESCs.
(C) Demethylation rate in WT and Tet TKO cells.
(D) Average CpG methylation level based on uncorrected hairpin sequencing counts for Tet TKO ESCs.
(E) Average nonCpG methylation level based on uncorrected hairpin sequencing counts for TKO ESCs.
(F) Relative difference in demethylation rate between WT and Tet TKO cells.
(G) Estimated, and conversion error corrected, 5mC and 5hmC distribution after HMM's application.
(H) HMM-derived maintenance methylation, *de novo* methylation and hydroxylation efficiencies.
(I) Integrative Genomics Viewer (IGV) snapshot across a gemoic region located at the Tet3 gene showing the distribution of CpGs across CpG-rich and -poor regions of RRHPoxBS.

Moreover, under primed conditions (serum/LIF) WT ESCs show a level of 78% methylation, where 56% of CpGs are fully methylated and 22% are found in a hemimethylated state (Figure 2A). 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).

These results agree with previously published whole genome methylation profiles (Ficz et al., 2013) that originate from the very same ESC sample. In addition, RRHPoxBS data show that hemimethylation is equally distributed among both DNA strands for all time points. Finally, we note that oxBS samples always display lower methylation levels than BS samples. This difference corresponds to the amount of 5hmC of each sample and it is mainly detected in the hemimethylated proportion indicating that a considerable amount of 5hmC might exist in a hemi(hydroxy)-methylated (5hmC/C or C/5hmC) state.

ESCs lacking Tet enzymes (Tet TKO) show only a marginal increase of methylated CpG dyads, i.e., 82% fully or hemimethylated, in comparison to WT under primed serum conditions. However, TKO cells show in relation to WT a clearly higher frequency of fully methylated CpGs (72%) and a reduced proportion of hemimethylated CpGs (hemiCpGs) (10%) (Figure 2D), which leads to the conclusion that in WT ESCs the enhanced presence of hemiCpGs is directly coupled to 5mC oxidation by Tets. All together, the above findings show that the presence of Tets has a considerable influence on DNA demethylation kinetics.

Tets influence nonCpG methylation levels

In addition, RRHPoxBS sequencing allowed us to accurately determine the amount, location and distribution of nonCpG methylation in WT and Tet TKO ESCs. For our analysis we considered only nonCpG positions which (1) are methylated above the conversion error, (2) show at least three methylated reads and (3) have a coverage of ≥ 10. In both WT and TKO we find CpA to be the most frequent methylated nonCpG motif (Figure S2A). Over time nonCpG methylation becomes gradually reduced upon cultivation in 2i. In WT ESCs the number of methylated nonCpGs was identical in BS and oxBS libraries, indicating that nonCpGs are not a substrate for Tet oxidation (Figure 2B). In Tet TKO cells the number of methylated nonCpGs is approximately doubled as compared with WT ESCs. Since nonCpG methylation is strictly dependent on the presence of *de novo* methylation activities by Dnmt3a/b (Arand et al., 2012), the higher nonCpG methylation in TKO cells both under primed (=2%) and naive (=0.6% after 168 h 2i) conditions (Figure 2E) points clearly toward an increased *de novo* methylation activity by Dnmt3a/b in the absence of Tet enzymes.

Tets affect *de novo* and maintenance machinery

The HMM-predicted uncorrected methylation levels for WT and Tet TKO ESCs fit well to the hairpin methylation data (Figure S3D), indicating a high prediction accuracy of our model. The HMM (Figure 3A) estimates a notable amount of 5hmC at all time points in WT ESCs. In Figure 2G the displayed amount of 5hmC refers to the sum of all possible 5hmC states. Our model estimates the majority of 5hmC to appear either paired with C (5hmC/C, C/5hmC) or 5mC (5hmC/5mC, 5mC/5hmC), while only a small proportion, mainly in regulatory regions, is present in a 5hmC/5hmC symmetric state (Figure 4B). The highest amount of 5hmC is observed at d3 meaning that WT ESCs display a transient increase of 5hmC after cultivation in 2i.

The enzymatic efficiencies estimated by GwEEP (Figure 2H) illustrate a mean maintenance methylation of about 61.4% at d0, which remains almost constant over time (60.1% at d6). 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. These estimations are 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 (Ficz et al., 2013; von Meyenn et al., 2016). In Tet TKO cells maintenance efficiency (58.8% at d0) lies at very similar levels with WT ESCs and remains stable over time (58.6% at d7) as well. On the other hand, *de novo* methylation efficiency exhibits the most pronounced difference between WT and Tet TKO cells. More specifically in Tet TKO *de novo* methylation efficiency begins from 20.3% at d0 and exhibits only a slight decrease over time (16.2% at d7). This prediction of our model is nicely substantiated by the elevated nonCpG methylation levels observed in Tet TKO data.

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

Redistribution of 5mC in the absence of Tets

Next, we 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 (Burger et al., 2013) 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 ESCs cultivated under serum/Lif conditions (Ficz et al., 2013) on the identical cell batch used for our study. We found that the majority of the WT ESC genome (58.1%) consists of large HMRs and PMDs (38.9%), while short LMRs and UMRs account for 0.02% and 2.8%, respectively (Figures 4A and 4C). The estimated methylation levels (sum of 5mC and 5hmC) for WT ESCs by our model fully agreed with those derived from WGBS (Figure 4E). 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 assigned 5mC and 5hmC modification levels, their distribution, and the corresponding Dnmt/Tet enzyme efficiencies determined by our model to CpGs of individual segments (Figures 4B, 4D, and 4F). All segments show a more pronounced loss of DNA methylation in WT compared with Tet TKO cells, where a higher frequency of fully methylated CpGs is retained across all time points (Figure 4D). Moreover, in HMRs and PMDs we observe a transient increase of 5hmC and hemiCpGs in WT ESCs, while the frequency of hemiCpGs in Tet TKO remains almost constant between d3 and d6 within all segments (Figures 4B and 4D). In contrast, WT ESCs exhibit a constant decrease in 5hmC and hemiCpGs in LMRs and UMRs over time. The increase of hemiCpGs in HMRs and PMDs (WT and

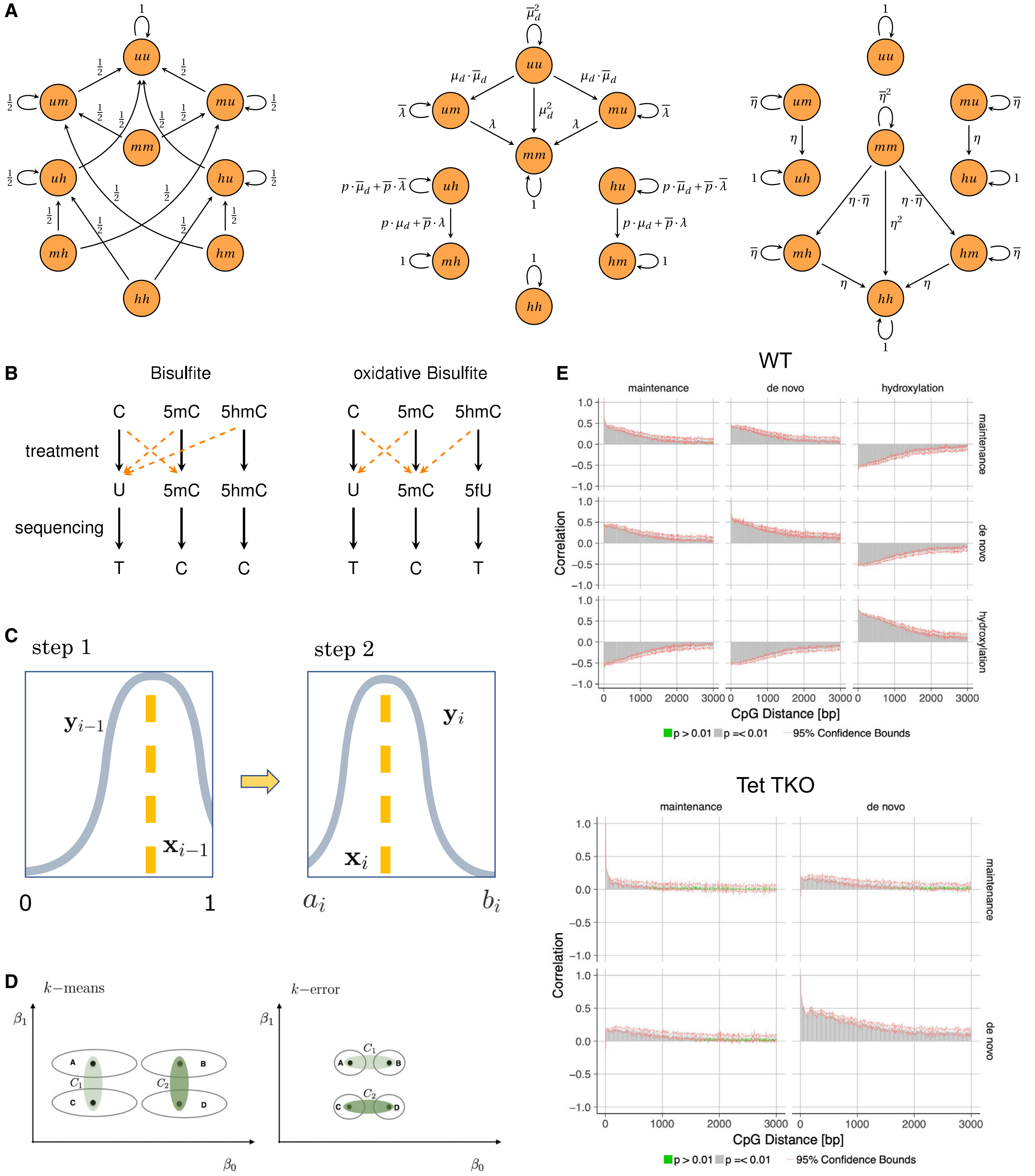


Figure 3. Model and computational methods

(A) Transitions between methylation states of a single CpG dyad; *u* indicates an unmethylated, *m* a methylated, and *h* a hydroxylated state of a CpG. μ_d describes the efficiency of *de novo* methylation, μ_m the efficiency of maintenance methylation and η the hydroxylation efficiency. λ represents the overall methylation efficiency (maintenance + *de novo*) that is defined as $\lambda = \mu_m + \mu_d - \mu_m \cdot \mu_d$. The parameter *p* describes the probability that maintenance methylation does not consider hemihydroxylated sites.

(B) Possible conversion errors during bisulfite and oxidative bisulfite sequencing.

(legend continued on next page)

TKO) is a clear sign of impaired maintenance methylation in naive ESCs linked to both the reported temporal increase in Tet expression and loss of Dnmt1 activity (Ficz et al., 2013; von Meyenn et al., 2016). Surprisingly, in HMRs and PMDs Tet TKO cells show a higher number of unmethylated CpGs as compared with WT ESCs under primed conditions (d0).

Tets prevent the spreading of DNA methylation

As expected, our model predicts high maintenance methylation efficiency in HMRs (69%) and PMDs (61%), but low maintenance efficiency in LMRs (32%) and UMRs (26%) (Figure 4F). In addition, we observe a relatively high *de novo* methylation efficiency at HMRs (18%) in primed ESCs. *De novo* methylation efficiency is negligible in UMRs and LMRs and clearly decreases upon cultivation in 2i for HMRs and PMDs. Finally, hydroxylation efficiency is high in UMRs (63%) and LMRs (55%), but low in HMRs (13%) and PMDs (24%). In Tet TKO, 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, 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.

An independently performed *k*-error clustering analysis (Figure 3D) returns two separate clusters and corroborates the hypothesis of distinct regions of opposing enzymatic activity (Figure S4C). Cluster 1, which displays high maintenance and low hydroxylation activity, contains the majority of CpGs assigned to HMRs and PMDs, while cluster 2 displays the opposite enzymatic profile and contains the majority of CpGs located in LMRs and UMRs (Table S4). To further quantify the opposing relationship between Dnmts and Tets we derived a spatial correlation measure between the efficiencies across the whole genome (Figures 3E and S5). Interestingly, in agreement with its apparent misregulation under the absence of Tets, maintenance autocorrelation almost disappears in Tet TKO cells, while the activity area of *de novo* methylation seems not to be affected in Tet TKO. Together, these results indicate regional differences and a clear antagonistic behavior between Dnmts and Tets.

Tets regulate Dnmts at TSSs and TFBSs

The genome-wide antagonistic effects of Dnmts' and Tets' activity across segments and clusters prompted us to plot, using DeepTools (Ramırez et al., 2014), the enzymatic efficiencies of CpGs across genes, histone marks, and ChIP profiled transcription factor binding sites (TFBSs) to investigate regularities and general local dependencies. The predicted efficiency profiles for maintenance methylation, *de novo* methylation, and hydroxylation correspond nicely to previously published Uhrf1 (GEO: GSE77779), Dnmt3a/b (GEO: GSE57413) and Tet1 (GEO: GSM659799) ChIP profiles, respectively (Figure 5A). In WT cells the efficiencies across genes and TFBSs show once more an opposing behavior. At transcription start sites (TSSs) and TFBSs high hydroxylation efficiency is coupled to reduced maintenance and almost absent *de novo* efficiency, with the inverse behavior persisting also upon 2i cultivation. Notably, under primed conditions *de novo* methylation has a strong presence in the gene body which disappears only after the transition to 2i (Figure 5B).

In Tet TKO ESCs the TSS-associated drop in maintenance methylation is much less pronounced and almost absent at d6/d7. In addition, *de novo* methylation exhibits almost no reduction upon 2i cultivation and is clearly maintained across the gene body. Regulatory regions marked by Sox2, H3K4me3 and Tet1 enrichment show a strong hydroxylation activity in WT cells, which is once more 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 only maintenance methylation activity across these regions (Figure 5C).

DISCUSSION

In this study we provide a comprehensive approach for measuring genome-wide DNA methylation and epigenetic efficiency profiling. GwEEP allowed us to infer how the activity of Dnmts and Tets contribute to modify CpGs and nonCpGs across the genome in a functional context. It is important to note that technically our RRHPoxBS data resemble and corroborate the overall methylation dynamics observed by classical RRBS and WGBS (Ficz et al., 2013; von Meyenn et al., 2016). However, RRHPoxBS data provide three important novel features: (1) A genome-wide representation of up to 4 million CpGs distributed across the genome, (2) a precise determination of 5mC and 5hmC levels at a single CpG dyad, and (3) a precise mapping of hemimethylated states and positions of nonCpG methylation. Compared with previous models (Ginno et al., 2020; Äijö et al., 2016a, b; Qu et al., 2013) the combination of RRHPoxBS and HMM allows us to calculate accurate 5mC and 5hmC levels by considering the conversion errors through BS and oxBS and, furthermore, to simultaneously infer the genome-wide efficiencies of maintenance methylation, *de novo* methylation as well as hydroxylation efficiencies.

The overall evaluation of our RRHPoxBS data showed that in mouse ESCs, as described previously for somatic cells (Arand et al., 2012), 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 with 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, which is in agreement with recent *in vitro* studies (Hashimoto et al., 2012; Valinluck

(C) Metropolis-Hastings update step: Assuming each efficiency is a linear time function, each next value is sampled using two truncated normal distributions in two consecutive steps. Step 1: ample the intercept $\mathbf{y}_{i-1}$ from the truncated normal with mean $\mathbf{x}_{i-1}$ and bounds [0, 1]. Step 2: sample the gradient $\mathbf{y}_i$ from the truncated normal distribution with mean $\mathbf{x}_i$ and bounds $[a_i, b_i]$, which depend on the sampled intercept $\mathbf{y}_{i-1}$ of Step 1.

(D) Clustering of estimated enzymatic efficiency with intercept β_0 and gradient β_1 for CpGs A, B, C, D using *k*-means versus *k*-error algorithm.

(E) Spatial auto- and cross-correlations of maintenance methylation, *de novo methylation* and hydroxylation efficiencies over the whole genome.

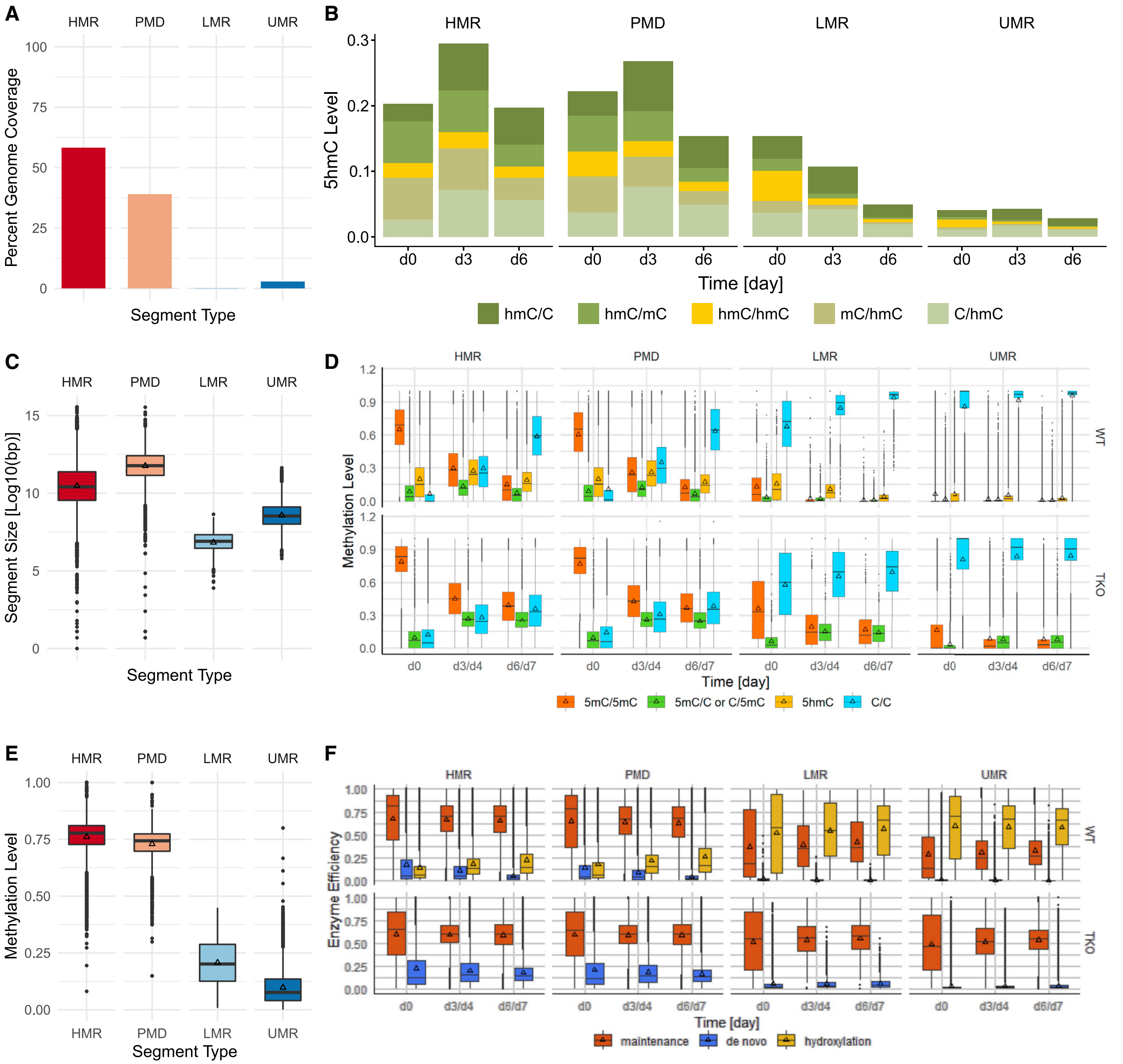


Figure 4. DNA methylation segmentation

(A) Percentages of HMRs, PMDs, LMRs, and UMRs across the genome based on WBGS data derived from WT ESC by Ficz et al. (2013).

(B) HMM estimated 5hmC distribution across the distinct segments based on RRHPoxBS of WT ESCs.

(C) Size distribution of the individual segment types based on WBGS data derived from WT ESCs by Ficz et al. (2013).

(D) HMM estimated methylation distribution of HMRs, PMDs, LMRs, and UMRs (methylation levels of the individual segment types for WT and Tet TKO ESCs).

(E) Methylation level of HMRs, PMDs, LMRs, and UMRs based on WBGS data derived from WT ESC by Ficz et al. (2013).

(F) Estimated HMM enzyme efficiencies in HMRs, PMDs, LMRs, and UMRs for WT and Tet TKO ESCs based on RRHPoxBS.

and Sowers, 2007). In contrast to equally distributed hemimethylation we observe a slight increase in the minus strand presence of nonCpG methylation. We cannot find a simple biological (sequence context) or technical (calling/mapping) explanation for this bias. NonCpG methylation is always occurring in close vicinity to CpG methylation (Arand et al., 2012), but in contrast to CpGs we find that nonCpGs are not a substrate for Tet enzymes, i.e., we do not find any indication of 5hmC in the nonCpG context. The amount of nonCpG methylation, however, is strongly enhanced in the absence of Tet enzymes, suggesting an increase of Dnmt3a and 3b efficiency in the absence of Tets. 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

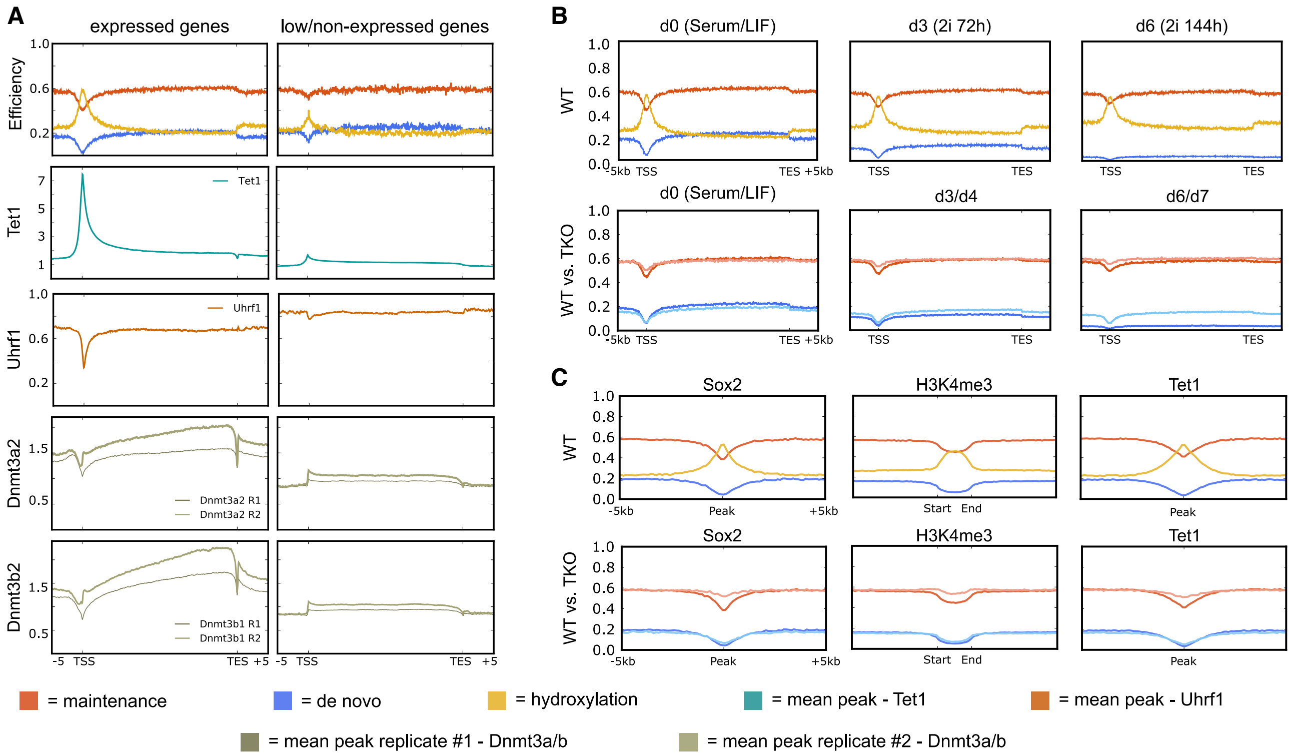


Figure 5. Enzymatic profiles across genes and TFBSs

(A) Comparison between ChIP profiles of epigenetic writers and estimated enzymatic efficiencies in WT ESCs at day 0 (serum/LIF) across expressed and low/non-expressed genes.

(B) Estimated enzymatic efficiencies in WT and Tet TKO ESCs across genes.

(C) Estimated enzymatic efficiencies of WT and Tet TKO ESCs at Sox2-, Tet1-, and H3K4me3-enriched regions obtained from ENCODE.

and *de novo methylation* efficiencies 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, TFBSs (Sox2, Pou5f1), and TSSs. 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) (Gu et al., 2018; Baubec et al., 2015).

The impairment of maintenance methylation has been identified so far as the main driver of 2i induced DNA demethylation (von Meyenn et al., 2016) and a role for Tet or oxidative cytosine forms, on the other hand, has only been recognized for selected loci (Ficz et al., 2013; von Meyenn et al., 2016). The comparison of WT and Tet TKO ESCs in this study, however, discloses a notable reduction within the demethylation rate of Tet TKO, compared with WT ESCs. On average, we detect a reduction in the demethylation rate of almost 50% from around 8%–4% loss per day. This indicates 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 ESCs compared with WT ESCs. Recently, a systematic investigation of genome-wide methylation profiles from various cell types carrying distinct Tet KO genotypes (López-Moyado et al., 2019) has detected, similar to our observations, a pronounced loss of DNA methylation in heterochromatic compartments, i.e., PMDs of Tet TKO mouse ESCs. Lopez-Moyado et al. propose a mutual exclusive localization of Dnmts and Tets in WT ESCs, while in Tet KO cells Dnmts invade domains that 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 ESCs we see an increase in maintenance methylation efficiency, but at the same time a reduction in HMRs and PMDs. In addition, Tet TKO cells exhibit a more stable, almost persistent, *de novo* methylation under naive conditions, which is further supported by the increased nonCpG methylation detected by RRHPoxBS and shows that in the absence of Tets ESCs fail to effectively downregulate *de novo* methylation in 2i. Taken together, these findings indicate a displacement of Dnmt1 and Dnmt3a/b, which fits to the hypothesized model by Lopez-Moyado et al.

Overall, we summarize that Tet enzymes work against methylation in three ways: (1) They guarantee an efficient conversion of

5 mC at accessible regions and act against its establishment during a cell replication, either via passive or active demethylation, (2) they inhibit the effectiveness of the maintenance machinery over regions that should remain unmethylated, and finally (3) they ensure an efficient downregulation of the *de novo* enzymes that cannot be observed in their absence.

Conclusions

We describe GwEEP, a 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. In GwEEP we generate strand-specific (hydroxy)methylation data and, using a sophisticated HMM, we infer the distribution of 5mC and 5hmC at individual CpGs across the genome and, furthermore, derive accurate efficiency profiles of Dnmts (*de novo* and maintenance) and Tets (hydroxylation). GwEEP also works with low amounts of DNA and is therefore suitable for demanding samples and rare cell types. Moreover, by combining our hairpin protocol with 5fC or 5caC detecting chemistry GwEEP is easily expandable for the estimation of 5fC/5caC distribution and the inference of formylation and carboxylation efficiencies of Tets. Our analysis of WT and Tet TKO mouse ESCs shows that Dnmts and Tets exhibit clear antagonistic efficiencies at individual CpGs. The comparison of WT and Tet TKO ESCs 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 within HMRs, guaranteeing the formation and maintenance of cell-type-specific methylation patterns.

STAR★METHODS

Detailed methods are provided in the online version of this paper and include the following:

- KEY RESOURCES TABLE
- RESOURCE AVAILABILITY
  - ○ Lead contact
  - ○ Materials availability
  - ○ Data and code availability
- EXPERIMENTAL MODEL AND SUBJECT DETAILS
- METHOD DETAILS
  - ○ Digestion of genomic DNA
  - ○ End-repair and A-tailing
  - ○ Ligation of hairpin-linker and sequencing adapter
  - ○ Enrichment of hairpin-DNA-adapter molecules
  - ○ BS and oxBS treatment
  - ○ Enrichment-PCR
  - ○ Sequencing
  - ○ Preparation of low-input-libraries
- QUANTIFICATION AND STATISTICAL ANALYSIS
  - ○ Processing sequencing data
  - ○ Hidden Markov modelling of single CpG methylation
  - ○ Bayesian inference for whole genome data
  - ○ Fit of whole genome data & uncertainty estimation
  - ○ *k*-error clustering
  - ○ Spatial correlation of enzymatic activity

SUPPLEMENTAL INFORMATION

Supplemental information can be found online at https://doi.org/10.1016/j.crmeth.2022.100187.

ACKNOWLEDGMENTS

We thank Jasmin Kirch and Gilles Gasparoni for performing sequencing using the Illumina HiSeq2500 system and Andreas Firczynski for the optimization of the processing pipeline for the hairpin sequencing data. Particular thanks to Jim Robinson and the entire Integrative Genomics Viewer (IGV) team for implementing the DSI-file format into the IGV genome browser. The project was funded by the DFG during the course of the SFB1309 - project number 325871075 and the German Epigenome Programme (DEEP) of the Federal Ministry of Education and Research in Germany (BMBF) (01KU1216 to J.W.). A.S. was supported by the German Federal Mindsistry of Research and Education grant for de.NBI (031L0101D).

AUTHOR CONTRIBUTIONS

Conceived and designed the experiments, C.K., P.G., W.R., V.W., and J.W.; wrote the manuscript, C.K., P.G., J.W., V.W., P.L.K., and J.Y.G.; performed the experiments, P.G., J.Y.G., J.A., F.v.M., and G.F.; processed the raw data, K.N.; designed/implemented the computational methods, C.K. and P.K.; analyzed the data, C.K.; performed meta analysis, C.K., P.G., K.N., A.S., and F.M.

DECLARATION OF INTERESTS

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

Received: November 28, 2021
Revised: January 19, 2022
Accepted: March 1, 2022
Published: March 21, 2022

REFERENCES

Äijö, T., Huang, Y., Mannerström, H., Chavez, L., Tsagaratou, A., Rao, A., and Lähdesmäki, H. (2016a). A probabilistic generative model for quantification of dna modifications enables analysis of demethylation pathways. Genome Biol. *17*, 1–22.

Äijö, T., Yue, X., Rao, A., and Lähdesmäki, H. (2016b). Luxglm: a probabilistic covariate model for quantification of dna methylation modifications with complex experimental designs. Bioinformatics *32*, i511–i519.

Arand, J., Spieler, D., Karius, T., Branco, M.R., Meilinger, D., Meissner, A., Jenuwein, T., Xu, G., Leonhardt, H., Wolf, V., et al. (2012). In vivo control of cpg and non-cpg dna methylation by dna methyltransferases. PLoS Genet. *8*, e1002750.

Arita, K., Ariyoshi, M., Tochio, H., Nakamura, Y., and Shirakawa, M. (2008). Recognition of hemi-methylated dna by the sra protein uhrf1 by a base-flipping mechanism. Nature *455*, 818–821.

Bachman, M., Uribe-Lewis, S., Yang, X., Williams, M., Murrell, A., and Balasubramanian, S. (2014). 5-hydroxymethylcytosine is a predominantly stable dna modification. Nat. Chem. *6*, 1049–1055.

Baubec, T., Colombo, D.F., Wirbelauer, C., Schmidt, J., Burger, L., Krebs, A.R., Akalin, A., and Schübeler, D. (2015). Genomic profiling of dna methyltransferases reveals a role for dnmt3b in genic methylation. Nature *520*, 243–247.

Bostick, M., Kim, J.K., Estève, P.O., Clark, A., Pradhan, S., and Jacobsen, S.E. (2007). Uhrf1 plays a role in maintaining dna methylation in mammalian cells. Science *317*, 1760–1764.

Bourc'his, D., and Bestor, T.H. (2004). Meiotic catastrophe and retrotransposon reactivation in male germ cells lacking dnmt3l. Nature *431*, 96–99.

Braunstein, S.L. (1992). How large a sample is needed for the maximum likelihood estimator to be approximately Gaussian? J. Phys. A: Math. Gen. *25*, 3813.

Burger, L., Gaidatzis, D., Schübeler, D., and Stadler, M.B. (2013). Identification of active regulatory regions from dna methylation data. Nucleic Acids Res. *41*, e155.

Charlton, J., Jung, E.J., Mattei, A.L., Bailly, N., Liao, J., Martin, E.J., Giesselmann, P., Brändl, B., Stamenova, E.K., Müller, F.J., et al. (2020). Tets compete with dnmt3 activity in pluripotent cells at thousands of methylated somatic enhancers. Nat. Genet. *52*, 819–827.

Chuang, L.S.H., Ian, H.I., Koh, T.W., Ng, H.H., Xu, G., and Li, B.F. (1997). Human dna-(cytosine-5) methyltransferase-pcna complex as a target for p21waf1. Science *277*, 1996–2000.

Ficz, G., Hore, T.A., Santos, F., Lee, H.J., Dean, W., Arand, J., Krueger, F., Oxley, D., Paul, Y.L., Walter, J., et al. (2013). Fgf signaling inhibition in escs drives rapid genome-wide demethylation to the epigenetic ground state of pluripotency. Cell Stem Cell *13*, 351–359.

Giehr, P., Kyriakopoulos, C., Ficz, G., Wolf, V., and Walter, J. (2016). The influence of hydroxylation on maintaining cpg methylation patterns: a hidden Markov model approach. PLoS Comput. Biol. *12*, e1004905.

Ginno, P.A., Gaidatzis, D., Feldmann, A., Hoerner, L., Imanci, D., Burger, L., Zilbermann, F., Peters, A.H., Edenhofer, F., Smallwood, S.A., et al. (2020). A genome-scale map of dna methylation turnover identifies site-specific dependencies of dnmt and tet activity. Nat. Commun. *11*, 1–16.

Globisch, D., Münzel, M., Müller, M., Michalakis, S., Wagner, M., Koch, S., Brückl, T., Biel, M., and Carell, T. (2010). Tissue distribution of 5-hydroxymethylcytosine and search for active demethylation intermediates. PLoS One *5*, e15367.

Goodstadt, L. (2010). Ruffus: a lightweight python library for computational pipelines. Bioinformatics *26*, 2778–2779.

Gu, T., Lin, X., Cullen, S.M., Luo, M., Jeong, M., Estecio, M., Shen, J., Hardikar, S., Sun, D., Su, J., et al. (2018). Dnmt3a and tet1 cooperate to regulate promoter epigenetic landscapes in mouse embryonic stem cells. Genome Biol. *19*, 1–15.

Habibi, E., Brinkman, A.B., Arand, J., Kroeze, L.I., Kerstens, H.H., Matarese, F., Lepikhov, K., Gut, M., Brun-Heath, I., Hubner, N.C., et al. (2013). Whole-genome bisulfite sequencing of two distinct interconvertible dna methylomes of mouse embryonic stem cells. Cell Stem Cell *13*, 360–369.

Hajkova, P., Jeffries, S.J., Lee, C., Miller, N., Jackson, S.P., and Surani, M.A. (2010). Genome-wide reprogramming in the mouse germ line entails the base excision repair pathway. Science *329*, 78–82. https://doi.org/10.1126/science.1187945. https://www.science.org/doi/abs/10.1126/science.1187945.

Hashimoto, H., Liu, Y., Upadhyay, A.K., Chang, Y., Howerton, S.B., Vertino, P.M., Zhang, X., and Cheng, X. (2012). Recognition and potential mechanisms for replication and erasure of cytosine hydroxymethylation. Nucleic Acids Res. *40*, 4841–4849.

He, Y.F., Li, B.Z., Li, Z., Liu, P., Wang, Y., Tang, Q., Ding, J., Jia, Y., Chen, Z., Li, L., et al. (2011). Tet-mediated formation of 5-carboxylcytosine and its excision by tdg in mammalian dna. Science *333*, 1303–1307.

Hermann, A., Goyal, R., and Jeltsch, A. (2004). The dnmt1 dna-(cytosine-c5)-methyltransferase methylates dna processively with high preference for hemimethylated target sites. J. Biol. Chem. *279*, 48350–48359.

Holliday, R., and Pugh, J.E. (1975). Dna modification mechanisms and gene activity during development. Science *187*, 226–232.

Ito, S., Shen, L., Dai, Q., Wu, S.C., Collins, L.B., Swenberg, J.A., He, C., and Zhang, Y. (2011). Tet proteins can convert 5-methylcytosine to 5-formylcytosine and 5-carboxylcytosine. Science *333*, 1300–1303.

Iyer, L.M., Tahiliani, M., Rao, A., and Aravind, L. (2009). Prediction of novel families of enzymes involved in oxidative and other complex modifications of bases in nucleic acids. Cell Cycle *8*, 1698–1710.

Ji, D., You, C., Wang, P., and Wang, Y. (2014). Effects of tet-induced oxidation products of 5-methylcytosine on dna replication in mammalian cells. Chem. Res. Toxicol. *27*, 1304–1309.

Kellinger, M.W., Song, C.X., Chong, J., Lu, X.Y., He, C., and Wang, D. (2012). 5-formylcytosine and 5-carboxylcytosine reduce the rate and substrate specificity of rna polymerase ii transcription. Nat. Struct. Mol. Biol. *19*, 831–833.

Kriaucionis, S., and Heintz, N. (2009). The nuclear dna base 5-hydroxymethylcytosine is present in purkinje neurons and the brain. Science *324*, 929–930.

Krueger, F. (2019). Babraham Bioinformatics: Trim Galore! (Babraham institute).

Krueger, Felix, Kreck, Benjamin, Franke, Andre, and Andrews, Simon R (2012). DNA methylome analysis using short bisulfite sequencing data. nature methods *9*, 145–151. https://doi.org/10.1038/nmeth.1828.

Kumar, M., and Patel, N.R. (2007). Clustering data with measurement errors. Comput. Stat. Data Anal. *51*, 6084–6101.

Kyriakopoulos, C., Giehr, P., and Wolf, V. (2017). H(o)ta: estimation of dna methylation and hydroxylation levels and efficiencies from time course data. Bioinformatics *33*, 1733–1734.

Leonhardt, H., Page, A.W., Weier, H.U., and Bestor, T.H. (1992). A targeting sequence directs dna methyltransferase to sites of dna replication in mammalian nuclei. Cell *71*, 865–873.

Li, E., Bestor, T.H., and Jaenisch, R. (1992). Targeted mutation of the DNA methyltransferase gene results in embryonic lethality. Cell *69*, 915–926.

Li, Heng, Handsaker, Bob, Wysoker, Alec, Fennell, Tim, Ruan, Jue, Homer, Nils, Marth, Gabor, Abecasis, Goncalo, and Durbin, Richard; 1000 Genome Project Data Processing Subgroup (2009). The Sequence Alignment/Map format and SAMtools. Bioinformatics *25* (*16*), 2078–2079. https://doi.org/10.1093/bioinformatics/btp352.

Liang, G., Chan, M.F., Tomigahara, Y., Tsai, Y.C., Gonzales, F.A., Li, E., Laird, P.W., and Jones, P.A. (2002). Cooperativity between dna methyltransferases in the maintenance methylation of repetitive elements. Mol. Cell. Biol. *22*, 480–491.

Lister, R., Pelizzola, M., Dowen, R., Hawkins, R., Hon, G., Tonti-Filippini, J., Nery, J., Lee, L., Ye, Z., Ngo, Q., et al. (2009). Human DNA methylomes at base resolution show widespread epigenomic differences. Nature *462*, 315–322.

Long, S.J., and Freese, J. (2006). Regression Models for Categorical Dependent Variables Using Stata (Stata press).

López-Moyado, I.F., Tsagaratou, A., Yuita, H., Seo, H., Delatte, B., Heinz, S., Benner, C., and Rao, A. (2019). Paradoxical association of tet loss of function with genome-wide dna hypomethylation. Proc. Natl. Acad. Sci. U S A *116*, 16933–16942.

Lorsbach, R., Moore, J., Mathew, S., Raimondi, S., Mukatira, S., and Downing, J. (2003). 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*, 637–641.

Maiti, A., and Drohat, A.C. (2011). Thymine dna glycosylase can rapidly excise 5-formylcytosine and 5-carboxylcytosine: potential implications for active demethylation of cpg sites. J. Biol. Chem. *286*, 35334–35338.

Marco-Sola, S., Sammeth, M., Guigó, R., and Ribeca, P. (2012). The gem mapper: fast, accurate and versatile alignment by filtration. Nat. Methods *9*, 1185–1188.

Martin, M. (2011). Cutadapt removes adapter sequences from high-throughput sequencing reads. Embnet. J. *17*, 10–12.

Martin-Herranz, D.E., Ribeiro, A.J., Krueger, F., Thornton, J.M., Reik, W., and Stubbs, T.M. (2017). currbs: simple and robust evaluation of enzyme combinations for reduced representation approaches. Nucleic Acids Res. *45*, 11559–11569.

Meilinger, D., Fellinger, K., Bultmann, S., Rothbauer, U., Bonapace, I.M., Klinkert, W.E., Spada, F., and Leonhardt, H. (2009). Np95 interacts with de novo dna methyltransferases, dnmt3a and dnmt3b, and mediates epigenetic silencing of the viral cmv promoter in embryonic stem cells. EMBO Rep. *10*, 1259–1264.

Meissner, A., Gnirke, A., Bell, G.W., Ramsahoye, B., Lander, E.S., and Jaenisch, R. (2005). Reduced representation bisulfite sequencing for comparative high-resolution dna methylation analysis. Nucleic Acids Res. *33*, 5868–5877.

von Meyenn, F., Iurlaro, M., Habibi, E., Liu, N.Q., Salehzadeh-Yazdi, A., Santos, F., Petrini, E., Milagre, I., Yu, M., Xie, Z., et al. (2016). Impairment of dna methylation maintenance is the main cause of global demethylation in naive embryonic stem cells. Mol. Cell. *62*, 848–861.

Okano, M., Bell, D.W., Haber, D.A., and Li, E. (1999). Dna methyltransferases dnmt3a and dnmt3b are essential for de novo methylation and mammalian development. Cell *99*, 247–257.

Okano, M., Xie, S., and Li, E. (1998). Cloning and characterization of a family of novel mammalian dna (cytosine-5) methyltransferases. Nat. Genet. *19*, 219–220.

Ono, R., Taki, T., Taketani, T., Taniwaki, M., Kobayashi, H., and Hayashi, Y. (2002). 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 Res. *62*, 4075–4080.

Oswald, J., Engemann, S., Lane, N., Mayer, W., Olek, A., Fundele, R., Dean, W., Reik, W., and Walter, J. (2000). Active demethylation of the paternal genome in the mouse zygote. Curr. Biol. *10*, 475–478.

Porter, J., Sun, M.A., Xie, H., and Zhang, L. (2015). Investigating bisulfite short-read mapping failure with hairpin bisulfite sequencing data. BMC genomics *16*, 1–9.

Qu, J., Zhou, M., Song, Q., Hong, E.E., and Smith, A.D. (2013). Mlml: consistent simultaneous estimates of dna methylation and hydroxymethylation. Bioinformatics *29*, 2645–2646.

Raiber, E.A., Murat, P., Chirgadze, D.Y., Beraldi, D., Luisi, B.F., and Balasubramanian, S. (2015). 5-formylcytosine alters the structure of the dna double helix. Nat. Struct. Mol. Biol. *22*, 44–49.

Ramırez, F., Dündar, F., Diehl, S., Grüning, B., and Manke, T. (2014). deeptools: a flexible platform for exploring deep-sequencing data. Nucleic Acids Res. *42*, 187–191.

Ramsahoye, B.H., Biniszkiewicz, D., Lyko, F., Clark, V., Bird, A.P., and Jaenisch, R. (2000). Non-cpg methylation is prevalent in embryonic stem cells and may be mediated by dna methyltransferase 3a. Proc. Natl. Acad. Sci. U S A *97*, 5237–5242.

Riggs, A.D. (1975). X inactivation, differentiation, and dna methylation. Cytogenet. Genome Res. *14*, 9–25.

Roberts, G.O., Gelman, A., and Gilks, W.R. (1997). Weak convergence and optimal scaling of random walk metropolis algorithms. Ann. Appl. Probab. *7*, 110–120.

Schoenberg, R. (1997). Constrained maximum likelihood. Comput. Econ. *10*, 251–266.

Sharif, J., Muto, M., Takebayashi, S.i., Suetake, I., Iwamatsu, A., Endo, T.A., Shinga, J., Mizutani-Koseki, Y., Toyoda, T., Okamura, K., et al. (2007). The sra protein np95 mediates epigenetic inheritance by recruiting dnmt1 to methylated dna. Nature *450*, 908–912.

Shen, D., and Lu, Z. (2006). Computation of Correlation Coefficient and its Confidence Interval in Sas (SUGI). Paper, 170–31.

Smith, Z.D., Chan, M.M., Mikkelsen, T.S., Gu, H., Gnirke, A., Regev, A., and Meissner, A. (2012). A unique regulatory phase of dna methylation in the early mammalian embryo. Nature *484*, 339–344.

Szwagierczak, A., Bultmann, S., Schmidt, C.S., Spada, F., and Leonhardt, H. (2010). Sensitive enzymatic quantification of 5-hydroxymethylcytosine in genomic dna. Nucleic Acids Res. *38*, e181.

Tahiliani, M., Koh, K.P., Shen, Y., Pastor, W.A., Bandukwala, H., Brudno, Y., Agarwal, S., Iyer, L.M., Liu, D.R., Aravind, L., et al. (2009). Conversion of 5-methylcytosine to 5-hydroxymethylcytosine in mammalian dna by mll partner tet1. Science *324*, 930–935.

Valinluck, V., and Sowers, L.C. (2007). Endogenous cytosine damage products alter the site selectivity of human dna maintenance methyltransferase dnmt1. Cancer Res. *67*, 946–950.

Walter, M., Teissandier, A., Pérez-Palacios, R., and Bourc'his, D. (2016). An epigenetic switch ensures transposon repression upon dynamic loss of dna methylation in embryonic stem cells. Elife *5*, e11418.

Ying, Qi-Long, Wray, Jason, Nichols, Jennifer, Batlle-Morera, Laura, Doble, Bradley, Woodgett, James, Cohen, Philip, and Smith, Austin (2008). The ground state of embryonic stem cell self-renewal. Nature volume *453*, 519–523. https://doi.org/10.1038/nature06968.

Ziller, M.J., Müller, F., Liao, J., Zhang, Y., Gu, H., Bock, C., Boyle, P., Epstein, C.B., Bernstein, B.E., Lengauer, T., et al. (2011). Genomic distribution and inter-sample variation of non-cpg methylation across human cell types. PLoS Genet. *7*, e1002389.

STAR★METHODS

KEY RESOURCES TABLE

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Chemicals, peptides, and recombinant proteins		
Acetonitrile	Fisher Chemical	A/0627/17
Ethanol	Fisher Chemical	E/0650DF/15
NaCl	Sigma Aldrich	S9888
NaOH	Sigma Aldrich	71690
TrisHCl	Roth	9090.3
EDTA	Grüssing	103051000
SDS	BioChemica	A2572
ATP	NEB	P0756S
dATP	Solis BioDyne	02-21-00100
dNTPs	Solis BioDyne	02-21-00100
CutSmart Buffer	NEB	B7204
HotStarTaq Buffer	Qiagen	203207
$MgCl_2$	Qiagen	203207
Dynabeads^{TM}-280 Streptavidin	ThermoFisher	11205D
AMPureXP® beads	Beckman Coulter	A63880
R.AluI	NEB	R0137
R.HaeIII	NEB	R0108
R.HpyCH4V	NEB	R0619
Klenow exo-	NEB	M0212
T4 DNA Ligase	NEB	M0202
HotStarTaq DNA Polymerase	Qiagen	203207
Critical commercial assays		
TrueMethyl®oxBS-Seq Module	TECAN	0414-32
Qubit^{TM}dsDNA HS Kit	Invitrogen	Q32851
High Sensitivity DNA Kit	Agilent	5067-4626
High Sensitivity DNA Reagents	Agilent	5067-4627
Deposited data		
RRHPBS – ESC, WT-Serum/LIF	This paper	GEO: GSM5176043
RRHPoxBS – ESC, WT Serum/LIF	This paper	GEO: GSM5176044
RRHPBS – ESC, WT 2i 72h	This paper	GEO: GSM5176045
RRHPoxBS – ESC, WT 2i 72h	This paper	GEO: GSM5176046
RRHPBS – ESC, WT 2i 144h	This paper	GEO: GSM5176047
RRHPoxBS – ESC, WT 2i 144h	This paper	GEO: GSM5176048
RRHPBS Tet_TKO Serum/LIF	This paper	GEO: GSM5176049
RRHPBS – ESC, Tet TKO 2i 48h	This paper	GEO: GSM5176050
RRHPBS – ESC, Tet TKO 2i 96h	This paper	GEO: GSM5176051
RRHPBS – ESC, Tet TKO 2i 168h	This paper	GEO: GSM5176052
RRHPBS – ESC, WT 2i 72h	This paper	GEO: GSM5694848
RRHPoxBS – ESC, WT 2i 72h	This paper	GEO: GSM5694849
Oligonucleotides		
Hairpin linker (HP) GGGCCATDD1DDATGGCCCT; 1 = dTbiotin; 5′-phosphate	This paper	NA
Sequencing adapter top (SA) A2A2T2TTT222TA2A2GA2G2T2TT22GAT2T; 2 = 5mC,	Illumina	PE Adapter

(*Continued on next page*)

Continued

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Sequencing adapter bottom (SA) GAT2GGAAGAG2A2A2GT2TGAA2T22AGT2A2; 2 = 5mC; 5′-phosphate	Illumina	PE Adapter
Enrichment primer forward (EnPrimerFor) AATGATACGGCGACCACCGAGATCTACACTCTTTCCCTACACGACGCTCTTCCGATCT	Illumina	TruSeq Universal Adapter
Enrichment primer reverse (EnPrimerRev) GATCGGAAGAGCACACGTCTGAACTCCAGTCAC[i5]ATCTCGTATGCCGTCTTCTGCTTG	Illumina	TruSeq Adapter, Index
Q1hmC TAMGATCAMGGCGAATMCGATMGAATCAMAGTGGCGMTTTAHGAAGTGCGAMAGCMTTAG; M = 5mC, H = 5hmC	Cambridge Epigenetix	Q1hmC
Q3hmC TAMGATCAMGGCGAATMHGATMGAATCMTTGTAGCGMTTTAHGAAGTGCGAMAGCHTTAG; M = 5mC, H = 5hmC	Cambridge Epigenetix	Q3hmC
Q6hmC TAMGATCAHGGCGAATGHHGATMGAATCAGTMAAGCGMTTTAHGAAGTGHGAMAGCHTTAG; M = 5mC, H = 5hmC	Cambridge Epigenetix	Q6hmC
QfC TACGATCAFGGCGAATCCGATCGAATCGTTTMGGCGCTTTACGAAGTGCGACAGCCTTAG;; M = 5mC, F = 5fC	Cambridge Epigenetix	QfC
QmC TAMGATMAMGGMGAATMMGATMGAATMTAGMTTGMGMTTTAMGAAGTGMGAMAGMMTTAG; M = 5mC	Cambridge Epigenetix	QmC
SQC TACGATCACGGCGAATCCGATCGAATCMAGATMGCGCTTTACGAAGTGCGACAGCCTTAG	Cambridge Epigenetix	SQC
Software and algorithms		
Trim Galore!	Krueger et al., 2012	https://www.bioinformatics.babraham.ac.uk/projects/trim_galore/
Cutadapt	Martin, 2011	https://github.com/marcelm/cutadapt
GEM-Mapper	Marco-Sola et al., 2012	https://github.com/smarco/gem3-mapper
Samtools	Li et al., 2009	http://www.htslib.org/
HPup	This paper, original code	https://zenodo.org/badge/latestdoi/429860269
GwEEP	This paper, original code	https://doi.org/10.5281/zenodo.6140150

RESOURCE AVAILABILITY

Lead contact

Further information and requests for resources and reagents should be directed to and will be fulfilled by the lead contact, Pascal Giehr (pgiehr@ethz.ch).

Materials availability

This study did not generate new unique reagents.

Data and code availability

- All data generated in this study as well as the HMM output has been deposited at NCBI's Gene Expression Omnibus (GEO) repository. Accession numbers are listed in the key resources table.

- All original code for RRHPoxBS sequencing data analysis (HPup) and the parallel implementation of the single CpG HMM suitable for a multi-core environment are shared via GitHub. DOIs are listed in the key resources table.
- Any additional information required to reanalyze the data reported in this paper is available from the lead contact upon request.

EXPERIMENTAL MODEL AND SUBJECT DETAILS

The DNA used for this study originate from Ficz et al. 2013. Briefly, ES cells were cultured without feeders either in standard serum-containing media (DMEM 4,500 mg/L glucose, 4 mM L-glutamine, 110 mg/L sodium pyruvate, 15% fetal bovine serum, 1 U/mL penicillin, 1 μg/mL streptomycin, 0.1 mM nonessential amino acids, 50 μM β-mercaptoethanol, and 103 U/mL LIF ESGRO) or under 2i culturing conditions (Ying et al., 2008) (serum-free N2B27 [Cat. DMEM/F12: GIBCO 21331; Neurobasal: GIBCO 21103; N2: Stem Cells SF-NS-01-005; B27: GIBCO 17504-044] supplemented with 103 U/mL LIF and Mek inhibitor PD0325901 [1 μM] and Gsk3β inhibitor CHIR99021 [3 μM]).

METHOD DETAILS

Digestion of genomic DNA

For GwEEP we use endo nucleases (RE) in order to enrich for selected genomic regions. This increases the sensitivity of the assay and, at the same time, reduces the sequencing costs considerably. The RE must not be sensitive against 5mC and 5hmC, in order to avoid a bias of the methylation analysis by blocked restriction. Ideally, the recognition/restriction site of the RE should not contain any CpG dyad. In the present protocol, we used a combination of three different REs, i.e., R.AluI, R.HaeIII and R.HpyCH4V to obtain reads from regulatory (HaeIII) as well as inter and intra genomic regions (AluI and HpyCH4V). For the customization of GwEEP we recommend cuRRBS, which determines the best suited RE(s) based on a list of *regions of interest* (Martin-Herranz et al., 2017).

Unless the amount of DNA is limited, we strongly advice a proper quality control using *Agarose-gel-elec-trophoresis* and *Qubit Fluorometer* based quantification. The best results are obtained by using 300–400 ng high quality DNA per used RE. However, as little as 6 ng per enzyme (18 ng in total) can be used. For each RE, prepare the reaction outlined in the table below and incubate for at least 3h at the temperature optimum of the RE (here 37°C), followed by a 20 min heat inactivation at 65°C or 80°C.

Digest of genomic DNA

Component	Concentration	Volume [μl]
Genomic DNA	6–400 ng	X
Endo Nuclease		
(R.AluI, R.HaeII or R.HpyCH4V)	10 U	Y
CutSmart Buffer	10x	2.0
ddH$_2$O	–	ad 20.0
Total Volume		20.0

Incubation over night might increase the sensitivity of the assay when using low DNA amounts or in case of sample impurities. This is only recommended for RE without star activity.

After inactivation the reactions are combined and subjected to a purification step using magnetic beads, i.e., SPRI® or AmpureXP® beads. Use a sample:bead ratio of 1:2 (60μL:120μL) for clean-up, wash twice with 200 μL freshly prepared 80% ethanol (EtOH) and elute the DNA in 17.5 μL of 1x CutSmart buffer.

At this stage we strongly recommend to add a suitable spike-in in order to calculate conversion rates of bisulfite and oxidative bisulfite reactions independently from the biological sample. We apply 4 pg of Sequencing Spike-in Control (CEGX) per sample (Table S1).

Addition of spike-in control

Component	Concentration [ng/μL]	Volume [μl]
Digested DNA	-	17.5
Sequencing Spike-in Control	0.008	0.5
Total Volume		18.0

End-repair and A-tailing

In the present study, we applied REs which generate blunt-end DNA fragments and therefore require a subsequent A-tailing. For this, we rely on the polymerase I large (Klenow) fragment, which lacks both $5' - > 3'$ and $3' - > 5'$ exonuclease activity (Klenow exo-). The reaction given in the table below is incubated for 30 min at 37°C and inactivated for 20 min at 75°C:

A-tailing reaction

Component	Concentration	Volume [μl]
Digested DNA	–	18.0
Klenow exo-	5 U/μL	1.0
dATP in 2x CutSmart	1 mM	1.0
Total Volume		20.0

When using REs producing sticky-ends, end-repair and A-tailing is required.

The A-tailing prevents self-ligation of DNA fragments and facilitates the ligation of hairpin linker (HP) and sequencing adapter (SA). Purification after heat-inactivation of Klenow exo- is not required. Instead, proceed immediately to section ligation of hairpin-linker and sequencing adapter.

Ligation of hairpin-linker and sequencing adapter

Prior to the ligation reaction, SA and HP have to be transferred into a double stranded state. First, solve each oligo nucleotide in 1xTE for a final concentration of 100μM. Next, join 10 μL SeqAdptTop and 10 μL SeqAdptBot in a 0.2 mL reaction tube and fill 20μL of HP in a second 0.2 mL reaction tube. Place both tubes into a thermocylcer and incubate for 5 min at 95°C followed by a slow cool-down of about 0.3°C/sec. This will facilitate the annealing of SeqAdptTop and SeqAdptBot as well as the proper folding of the HP (Figure S1A).

The HP comprises four distinct features, (i) a 3′ T overhang complementary to the A-overhang of the DNA molecules, (ii) unmethylated cytosine within the sequence, which permits the calculation of C-to-T conversion rate after sequencing, (iii) a unique molecular identifier (UMI), which allows to identify individual original ligation events and the removal of clonal PCR amplificates and (iv) a biotinylated T, for enrichment of HP containing DNA molecules after ligation. The ligation of SA and HP are catalysed by the T4 DNA ligase following the reaction outlined below.

Ligation reaction

Component	Concentration	Volume [μl]
A-tailed DNA	–	20.0
rCutSmart Buffer	10x	0.5
ATP	10 mM	2.5
Sequencing adapter (SA)	100 μM	0.5
Hairpin linker (HP)	100 μM	0.5
T4 DNA Ligase	2000 U/μL	1.0
Total Volume		25.0

The reaction is incubated at 16°C for 16 h and inactivated at 65°C for 10 min. The ligation of SA and HP is a stochastic event, meaning that three distinct types of DNA molecules will be generated: (i) molecules with SA on both ends, (ii) molecules with HP on both ends and (iii) molecules with SA on one side and HP on the other side.

Safe stopping point. DNA can be stored for 24 h at 4°C or long term (days to weeks) at −20°C or −80°C. Avoid repeated thawing and freezing, as this can lead to strand breaks and significantly reduces the number of usable DNA molecules.

Including other modifications such as 5mC, 5hmC, 5fC and 5caC, in the HP sequence allows to determine the conversion of non-canonical cytosine forms during BS and oxBS treatment.

Enrichment of hairpin-DNA-adapter molecules

In order to deplete the unwanted non-HP molecules (SA on both sides) the library is subjected to a purification step using streptavidin coated beads. Start with an AMPure beads clean-up to remove proteins and excessive HPs that would likely saturate the streptavidin beads. Use a sample:bead ratio of 1:2 (25μL:50μL), wash twice with freshly prepared 80% EtOH and elute in 50μL ddH_2O.

1. Per library, i.e., sample, transfer 10μL of Dynabeads^{TM} M-280 Streptavidin into a 1.5 mL reaction tube. Place the tube onto a magnetic stand and carefully remove the supernatant without disturbing the beads.

2. Add 1 mL 1xBW buffer (5 mM TrisHCl, 0.5 mM EDTA, 1 M NaCl), vortex thoroughly, place the mixture back onto the magnetic stand and wait until the solution is clear. Carefully remove and discard the supernatant.
3. Repeat step (2) for a total of 2 wash steps. At the end of wash step three, solve the beads in 50μL 2xBW buffer per library.
4. Transfer 50 μL of beads to each library, mix briefly by flicking the tube and incubate for 20 min at room-temperature while rotating.
5. Collect all liquid at the bottom of the tube though brief centrifugation and place the tube onto the magnetic stand. Wait until the solution is clear and carefully remove and discard the supernatant.
6. Add 200 μL 0.1N NaOH and mix by vortexing. Spin down all liquid, place the tube back onto the magnetic stand and wait until the solution is clear. Remove and discard the supernatant.
7. Repeat step (6) for a total of 2 wash steps.
8. Add 200 μL 0.5xTE buffer to the beads, mix by vortexing and place the tube back onto the magnetic stand. Wait until the solution is clear. Remove and discard the supernatant.
9. Repeat step (8) for a total of 2 wash steps. At the end of the second wash step, remove TE buffer completely, but do not let the beads dry fully.
10. Resuspend the beads in 50 μL 1xTE with 1% SDS.
11. Lock the tube and incubate the mixture for 5 min at 100°C in order to dissolve the the biotin-streptavidin-interaction. Place on ice for 2 min.
12. Briefly spin down and place the tube onto the magnetic stand. Wait until the solution is clear and transfer the supernatant containing the DNA into a new 1.5 mL DNA low-bind reaction tube.

Using 1.5 mL DNA-low-bind reactions tubes might increase the final yield of the library. This relevant when working with low input samples.

Safe stopping point. DNA can be stored for 24 h at 4°C or long term (days to weeks) at −20°C or −80°C. Avoid repeated thawing and freezing, as this can lead to strand breaks and significantly reduces the number of usable DNA molecules.

BS and oxBS treatment

For bisulfite and oxidative bisulfite treatment, we rely on the TrueMethyl®oxBS Module from TECAN. Here, we will provide a short summary of the individual steps of the protocol.

DNA purification

Bring the Magnetic Bead Solution 1 (TrueMethyl Kit) and the Binding Buffer 1 (TrueMethyl Kit) to room temperature. Mix bead and buffer according to manufacturer's instruction. To 50 μL DNA (in 1x TE and 1% SDS), add 100 μL bead-buffer solution and mix well by pipetting. Incubate at room temperature for 20 min. Transfer the solution onto the magnet and wait until the solution is clear (5 min). Carefully remove and discard the supernatant and wash 3x with 200 μL freshly prepared 80% acetonitrile. Let the beads dry for 5 min until all acetonitrile has evaporated. Remove the reaction tube from the magnet and solve the beads in 18 μL denaturation solution (TruMethyl Kit). Elute and denature the DNA at 37° C for 5 min. Put the tube back onto the magnet and wait until the solution is clear.

DNA oxidation

Transfer 9 μL into two 0.2 mL reaction tubes, respectively. To the tube intended for BS, add 1 μL ultrapure water (TrueMethyl Kit) in the other, meant for oxBS, add 1 μL oxidant solution. Mix both reactions by vortexing and immediately incubate at 40°C for 10 min. Centrifuge at 14000 x g for 10 min. A black precipitate will form at the bottom of the tube of the oxBS reactions. The supernatant should remain orange. Any other color (yellow, brown, transparent) indicates impurity of the sample and in our experiments indicates failing of the 5hmC oxidation. Transfer the supernatant from oxBS samples into a new 0.2 mL reaction tube.

Bisulfite treatment

Add 700 μL of the Bisulfite Dilutent (TrueMethyl Kit) to one aliquot of Bisulfite Reagent (TrueMethyl Kit). Incubate for 15 min at 60°C while shaking. Spin down briefly and add 30 μL of the bisulfite solution to BS and oxBS sample, respectively. Incubate the reaction according to the temperature profile outlined in the table below.

Bisulfite conversion temperature profile

Step Number	Incubation Step	Temperature	Time
1	Denaturation	95°C	05:00 min
2	Sulfonation	60°C	20:00 min
3	Denaturation	95°C	05:00 min
4	Sulfonation	60°C	40:00 min
5	Denaturation	95°C	05:00 min
6	Sulfonation	95°C	45:00 min
7	Hold	20°C	≤ 16:00:00 h

Prepare Magnetic Bead Solution 2 by mixing 2.4 μL Magnetic Beads Solution (TrueMethyl Kit) with 200 μL Binding Buffer 2 (TrueMEthyl Kit). Mix well by vortexing. Add 160 μL bead solution to 40 μL BS and 40 μL oxBS reaction, respectively. Mix well by pipetting and incubate for 5 min. Place the reaction onto an magnetic stand, wait until the solution is clear and discard the supernatant. Wash the beads with 200 μL freshly prepared 70% ethanol and resuspend the beads in 200 μL Desulfonation Buffer (TrueMethyl Kit). Incubate for 5 min, remove the supernatant and wash the beads twice with 200 μL 70% ethanol. Let the beads dry completely for 15 min and elute the DNA in 12.5 μL Elution Buffer (TrueMethyl Kit).

Enrichment-PCR

The enrichment PCR amplfies the library molecules and at the same time introduces the remaining part of the sequencing adapter, i.e., indices (i5 and i7) as well as the sequences required for flow cell binding (P5 and P7). The following table depicts the reaction conditions and the temperature profile of the enrichment PCR, respectively.

Safe stopping point. DNA can be stored for 24 h at 4°C or long term (days to weeks) at −20°C or −80°C. Avoid repeated thawing and freezing, as this can lead to strand breaks and significantly reduces the number of usable DNA molecules.

Enrichment PCR

Component	Conc.	Vol. [μl]
BS or oxBS DNA	–	10.0
HotStarTaq Buffer	10x	5.0
$MgCl_2$	25 mM	2.0
dNTPs	10 mM	4.0
EnPrimerFor	10 μM	0.8
EnPrimerRev	10 μM	0.8
HoStarTaq DNA Polymerase	5 U/μL	0.7
ddH_2O	–	26.7
Total Volume		50.0

Enrichment PCR temperature profile

PCR Step	Temp.	Time	# of Cycles
Initial Denaturation	95°C	5:00 min	1
Denaturation	94°C	0:30 min	
Annealing	58°C	1:00 min	7–18
Elongation	72°C	1:00 min	
Final Elongation	72°C	7:00 min	1
Hold	4°C	∞	1

After amplification, libraries are purified using AMPureXP®beads with a sample:bead ratio of 1:1 (50μL:50μL), wash twice with freshly prepared 80% EtOH and elute in 10μL ddH_2O or 0.1xTE. QC is performed by quantification of 2 μL of the library using Qubit HS Fluorometer and fragment size distribution is determined by loading 1 μL using the Agilent Bioanalyzer HS assay. In our case, the average library size lies by about 400bp. The library size might vary depending on the used restriction enzymes. A typical Bioanalyzer profile of RRHPoxBS libraries is given in Figure S1D.

Sequencing

The library is suited for any Illumina sequencing device and does not require custom sequencing primer. For larger genomes e.g. human or mouse, we recommend sequencing on an Illumina NextSeq, HiSeq or NovaSeq platform using ≥ 2x100bp paired-end mode. Note that longer reads allow more frequent detection of the HP and thus more accurate estimation of the duplication rate (Table S1). The indicated enzyme combination requires sequencing of 40–50 million reads to obtain an average coverage of about 10x. We recommend using 10–20% PhiX as a spike-in for the final library pool. See also Illumina guidelines for sequencing of bisulfite libraries, i.e., low complexity libraries.

Preparation of low-input-libraries

For the generation of Hairpin-RRBS libraries using 18 ng of genomic DNA, we performed three individual restriction reactions (R.AluI, R.HaeIII and R.HpyCH4V) in 1x CutSmart buffer with 6 ng of mouse ES cell DNA (72h 2i) each. 5 U of restriction enzyme were used in a total reaction volume of 10 μL. Reactions were incubated at 37°C for 3 h. After heat inactivation at 80°C for 20 min, reactions were pooled and 0.5 μL of CEGX spike-in control (1 pg/μL) were added. A-tailing was performed by adding 1 μL of 10X CutSmart buffer, 1 μL of 1 mM dATP, 2.4 μL of 5 U/μL Klenow exo- (NEB M0212S) and 5.1 μL ddH2O for a total reaction volume of 40 μL. Reaction was incubated at 37°C for 30 min and heat inactivated at 75°C for 20 min. For ligation of HP and SA, 0.5 μL of 100 μM HP, 0.5 μL of 100 μM SA, 5 μL of 10 mM ATP and 2 μL of 2000 U/μL T4 DNA Ligase (NEB M0202T) were added for a total volume of 48 μL, ligation reaction was incubated at 16°C over night. Depletion of unbound HP using AMPure XP beads, enrichment of HP-DNA using Dynabeads, oxidation reaction and bisulfite conversion (both according to TECAN Methyl® oxBS Module manual), enrichment PCR and QC were done similarly to our protocol outlined in the above methods section. Sequencing was carried out in a 2x 150 bp mode on the Illumina MiSeq using MiSeq Reagent Nano Kit v2 (300-cycles).

QUANTIFICATION AND STATISTICAL ANALYSIS

Processing sequencing data

In a first instance, the sequenced reads are subjected to a hairpin sequencing pipeline (HPup) that will restore the double strand information and output the methylation counts (uncorrected) for CpGs, GpCs and CpHs dyads. For this, the sequencing data was processed similarly to what has been previously described by Porter *et al.* (Porter et al., 2015). We used the Ruffus pipeline framework (Goodstadt, 2010) for the implementation of our workflow. In an initial step sequencing adapter and low quality base calls ($Q < 20$) were trimmed from the ends of each read (FastQ files) using TrimGalore! (Krueger, 2019). Next, all reads were screened and purged from the hairpin linker sequence using cutadapt (Martin, 2011). The hairpin linker sequence is stored in an additional file and used to calculate the C-to-T conversion rate during (oxidative) bisulfite treatment and furthermore, can be used to identify redundant reads generated by PCR.

Trimmed read pairs (read 1 and read 2 from paired-end-sequencing) were locally aligned with the Smith-Waterman algorithm. We allow for G-to-A and C-to-T mismatches to cope with bisulfite treatment. After aligning the bisulfite sequence of the read pairs, we can reconstitute the genomic sequence, which allows for a faster, more efficient and precise mapping of the reads to the reference genome. In addition, the pipeline determines the methylation state of each cytosine and classifies symmetric dyads, i.e., CpGs and GpCs into fully-methylated (both DNA strands are methylated), hemi-methylated (only one strand is methylated) at the plus strand, hemi-methylated on the minus strand and unmethylated (both strands are unmethylated). The resulting sequences were aligned to the mouse genome (mm10) with GEM-mapper (beta build 1.376) (Marco-Sola et al., 2012), after which the methylation information was reintroduced with a custom pileup function based on HTSJDK and ratios for the four methylation states were calculated for each cytosine. Our pipeline is suited for read mapping against any reference genome, i.e., species.

The hairpin pipeline comes with a configuration file in which all required parameters can be defined by the user. This includes the path to the input/output directory and to the adapter and hairpin linker sequences for trimming. The library generates three output files per sample: (i) a log file, containing the information about the individual processing steps (parameter, success and duration of individual steps). (ii) a statistic file which provides the read count after each processing step (raw reads, trimmed, paired, aligned) as well as the conversion rates of C, 5mC, 5hmC and 5fC, calculated based on hairpin linker and spike-in sequence. Lastly, (iii) a DSI-file (double strand information), which stores the methylation information of each cytosine is generated.

The DSI-file is a header-less, tab-separated text-file and contains in total 11 columns. Columns 1 to 7 contain general information: (1) the chromosome number, (2) the genomic location of the analyzed cytosine, (3) the strand of the analyzed cytosine, (4) total number of reads, (5) number of methylated reads, (6) number of unmethylated reads and (7) sequence context of the cytosine (CG, GC or nonCpG). Columns 8 to 11 contain double strand specific methylation information: (8) counts of fully methylated dyads (CpG or GpC; NA = nonCpG), (9) counts of plus-strand methylated dyads, (10) counts of minus strand methylated dyads and (11) count of unmethylated dyads. DSI-files can be visualised in the Integrative Genomics Viewer (IGV). Each sample will be displayed as bar diagram-track, in which each cytosine is represented by a stacked bar summarizing the frequency of the methylation states (fully = orange, hemi-plus = dark-green, hemi-minus = light-green and unmethylated = blue). Hovering the mouse cursor above the bar will show the entire 11-column information of the DSI-file for the given position.

Hidden Markov modelling of single CpG methylation

We provide the DSI-files containing the BS and oxBS counts for each sequenced CpG as input to an HMM. Incorporating additionally the conversion rates of the measurement process (Figure 3B) we link the HMMs, describing the oxidative and the non-oxidative hairpin bisulfite sequencing, to accurately determine (hydroxy)methylation levels and the efficiencies of the involved enzymes over time.

The computational core of our estimation is the model previously described in (Giehr et al., 2016; Kyriakopoulos et al., 2017). This HMM combines the processes of cell division, i.e., DNA replication, methylation (*de novo* and maintenance), as well as hydroxylation (Figure 3A). In addition, given the BS and oxBS counts of a single CpG we use here a Bayesian inference framework (Figure 3C) to estimate the distribution of methylation states and to determine the posterior distribution of the corresponding methylation efficiencies.

The advantage of our Bayesian inference scheme compared to alternative approaches such as Maximum Likelihood Estimation is that it yields useful estimates also in case of relatively small read counts. We apply the Metropolis-Hastings (MH) algorithm to sample from the posterior distribution of the Dnmts' and Tets' efficiencies that best explain the DNA methylation levels of a single CpG at three available time points. Finally, our model also provides an estimate for the maintenance of 5hmC during replication, i.e., the probability p that a given 5hmC is *not* recognized by Dnmt1 as a "methylated" state after replication.

Hidden and observable states

The hidden states of the model, $\mathcal{S} = \{u, m, h\}^2$, correspond to the different modifications, e.g. the cytosines (C) on both strands are unmethylated or the C on the upper strand is methylated, while the C on the lower strand is unmethylated, etc. The observable states, $\mathcal{S}_{\text{obs}} = \{\text{T}, \text{C}\}^2$, are those that we measure after bisulfite (BS) or oxidative bisulfite (oxBS) hairpin sequencing. Let the vector $\pi(t)$ be the hidden states distribution at time t and let $\pi(i,t) = P(\mathcal{X}(t) = i)$ represent the entry of $\pi(t)$ that corresponds to state $i \in S$. The transition matrix of the hidden states is defined as $\mathbf{P}(t) = \mathbf{D}(t) \cdot \mathbf{M}(t) \cdot \mathbf{H}(t)$, where $\mathbf{D}(t)$ describes the modifications due to cell division, $\mathbf{M}(t)$ the modifications due to methylation, and $\mathbf{H}(t)$ the modifications due to hydroxylation.

Note that for $\mathbf{D}(t)$ we can omit the time parameter t since it is time-independent, while the other two matrices depend on t. Note also that the HMMs of BS and oxBS experiments have both the same distribution $\pi(t)$ for the hidden states (as for both experiments the same cell population is used), but different emission probabilities and that $\pi(t)$ is given by $\pi(t) = \pi(0) \cdot \prod_{k=1}^{t} \mathbf{P}(k)$. Let the vectors $\pi_{bs}(t), \pi_{ox}(t)$ be the observable states distribution at time t, with entries $\pi_{bs}(j,t)$ and $\pi_{ox}(j, t)$, $j \in \mathcal{S}_{obs}$, for the BS and oxBS experiments, respectively. We get then $\pi_{bs}(t) = \pi(t) \cdot \mathbf{E}_{bs}(t)$ and $\pi_{ox}(t) = \pi(t) \cdot \mathbf{E}_{ox}(t)$, where the entries of the emission matrices $\mathbf{E}_{bs}(t)$ and $\mathbf{E}_{ox}(t)$ are given in Table S2.

Initial distribution

Let $n_{bs}(j,t)$ and $n_{ox}(j,t)$ be the number of times that state $j \in \mathcal{S}_{obs}$ has been observed during independent hairpin bisulfite (BS) and oxidative hairpin bisulfite (oxBS) measurements out of a certain number of reads (mean coverage of all samples $\approx$ 20x) at time t. Since we assume that $t = 0$ is the time of the first measurement, we have observations at $t = 0$ and can estimate the unknown initial distribution over the hidden states using maximum likelihood estimation (MLE). For this, we have to solve the optimization problem: $\pi(0)^* = arg\ max_{\pi(0)} \mathcal{L}_1(\pi(0))$, subject to the constraint $\sum\limits_{i \in \mathcal{S}} \pi(i,0) = 1$, where $\mathcal{L}_1(\pi(0)) = \prod\limits_{j \in \mathcal{S}_{obs}} \pi_{bs}(j,0)^{n_{bs}(j,0)} \cdot \pi_{ox}(j,0)^{n_{ox}(j,0)}$.

Estimation of the efficiencies

Let $\mathbf{v} = (\beta_0^{\mu_m}, \beta_1^{\mu_m}, \beta_0^{\mu_d}, \beta_1^{\mu_d}, \beta_0^{\eta}, \beta_1^{\eta}, p) \in \mathbb{R}^v$, be the vector of seven, i.e., $v = 7$, unknown parameters. We assume here that the efficiencies are linear functions of time (except for p) and so $\mathbf{v}$ contains the coefficients of these functions, e.g., $\mu_m(t) = \beta_0^{\mu_m} + t \cdot \beta_1^{\mu_m}$. After determining $\pi(0)$ we define the likelihood

$$\mathcal{L}_2(\mathbf{v}) = \prod_{t \in T_{obs} \setminus \{0\}} \prod_{j \in \mathcal{S}_{obs}} \pi_{bs}(j,t)^{n_{bs}(j,t)} \cdot \pi_{ox}(j,t)^{n_{ox}(j,t)}, \qquad \text{(Equation 1)}$$

assuming that the cells divide every 24 hours, t ranges over all days at which measurements were made after d0, and that all observations are independent. The independence assumption is well justified since during the measurement only a very small fraction of cells is taken out of a large pool and hence it is unlikely that we pick two cells with a common descendant. Since the efficiencies are probabilities we have the constraint that for all time points in T_{obs} and all efficiencies we have $0 \le \beta_0 + \beta_1 \cdot t \le 1$ and $0 \le p \le 1$. Finally, considering the forward Kolmogorov equation for the HMM and the first and second partial derivatives w.r.t. $\mathbf{v}$ we can efficiently estimate $\mathbf{v}^* = \text{argmax}_{\mathbf{v}} \mathcal{L}_2(\mathbf{v})$ as well as confidence intervals for a MLE, as has been shown in detail in (Giehr et al., 2016).

Bayesian inference for whole genome data

After QC there is available double stranded single base pair resolution data from BS and oxBS for $3,022,903$ CpGs in WT cells and for $3,151,985$ CpGs from BS data in Tet TKO cells (Figure S3B). In case of each of $1,464,801$ CpGs in WT and of $1,352,297$ in Tet TKO with only one or two observation time points available we predict only (hydroxy)methlation levels by performing a MLE as described in the initial distribution section. In case of a CpG with three observation time points ($1,558,102$ in WT and $1,799,688$ in Tet TKO) we apply the HMM for estimating the (hydroxy)methylation efficiencies. Using a computer cluster consisting of 32 machines with 16 physical kernels each, we are able to efficiently parallelize the computations for large batches of CpGs.

Due to the low depth sequencing per time point and experiment (40*x* for BS, 29*x* for oxBS in WT, and 14*x* coverage for BS in Tet TKO on average) we expect the asymptotic properties of the MLE around the true parameter value not to hold (Braunstein, 1992; Long and Freese, 2006), especially close to the boundary constraints (Schoenberg, 1997). For this reason, we use a Bayesian Inference (BI) approach to get the posterior distribution of the model parameters, i.e., the efficiencies over time. For all CpGs we choose as prior distribution the multivariate normal distribution $\mathcal{N}(\boldsymbol{\mu}, \Sigma)$, where the mean $\boldsymbol{\mu}$ is the average of the estimated efficiencies in (Giehr et al., 2016) and Σ is the average of the corresponding covariance matrices.

Metropolis-Hastings

We apply BI by sampling from the multi-dimensional posterior $P(\mathbf{v}|\text{data}) = \frac{\mathcal{L}_2(\text{data}|\mathbf{v})P(\mathbf{v})}{\int_{\mathbf{v}} P(\text{data},\mathbf{v})}$. To avoid approximating the normalizing factor $\int\limits_{\mathbf{v}} P(\text{data}, \mathbf{v})$, we make use of a Metropolis-Hastings MCMC approach using an asymmetric and truncated proposal distribution. The bounds of the truncation are determined s.t. the constraints for the efficiencies constantly hold for the time span of the observations, i.e., efficiencies are in $[0,1]$ for all $t \in [0, t_{\text{max}}]$. Hence, in every state $\mathbf{x} \in \mathbb{R}^v$ of the MCMC we generate the next sample from a product of truncated univariate normals $\mathcal{N}(\mathbf{y}) = \prod\limits_i f(\mathbf{y}_i | \mathbf{x}_i, \sigma_i^2 / c, a_i, b_i)$, around the current MCMC point $\mathbf{x}$, where $\mathbf{x}_i$ refers to the $i-$ th entry of the parameter vector for $i = 1,\ldots,7$, σ_i^2/c is the univariate normal variance and a_i, b_i are the truncation bounds for parameter $\mathbf{x}_i$. Consider position i where $\mathbf{y}_i$ refers to the gradient of an efficiency and $\mathbf{y}_{i-1}$ to the corresponding intercept. We sample the next value for each efficiency by sampling first the intercept $\mathbf{y}_{i-1}$ value from the truncated normal distribution within the interval $[a_{i-1}, b_{i-1}] = [0, 1]$ and based on this realization we sample the gradient $\mathbf{y}_i$ value from the truncated normal in $[a_i, b_i]$, where $a_i = -\mathbf{y}_{i-1}/t_{\max}$, $b_i = (1 - \mathbf{y}_{i-1})/t_{\max}$ (Figure 3C). The bounds of probability p are set as those of an intercept, i.e., $[a_i, b_i] = [0, 1]$.

Note that the variance of parameter $\mathbf{x}_i$ we used for the proposal distribution is the same as the variance of the prior distribution $\sigma_i^2 = \Sigma_{i,i}$ normalized by a scale factor c. Because the efficiency of Metropolis-Hastings algorithm crucially depends on the scaling of the proposal density, we empirically choose a $c = 50$ to normalize the standard deviation of the proposal distribution s.t. the average MCMC acceptance ratio is around 25% of the total number of generated samples (Roberts et al., 1997). As final estimators of the BI method we get the sample mean of the posterior distribution and we build credible intervals using the corresponding sample covariance.

Fit of whole genome data & uncertainty estimation

We compare the levels of CC, TT and CT-TC CpG dyads for the whole genome data in WT and Tet TKO experiments with the probabilities of the observable states predicted by the two HMMs using MLE or BI (Figures S3C and S3D). To quantify the goodness of fit we computed the average Kullback-Leibler divergence $D_{KL}(P||Q) = \sum_i P(i) \ln \frac{P(i)}{Q(i)}$ between the data distribution P and the distribution Q predicted by the model (Table S3). Even though the average KL divergence between the data and the model is smaller for MLE than for BI for both WT and Tet TKO, the uncertainty around the ML estimates is much higher. We quantified this by computing the volume of the hyper-ellipse of a multivariate-normal distribution

$$V = \frac{2\pi^{v/2}}{v\Gamma(v/2)} \left(\chi^2_{\text{crit}}\right)^{v/2} |\Sigma_{\mathbf{v}^*}|^{1/2},$$

Where v is the number of parameters, $|\Sigma_{\mathbf{v}^*}|$ is the determinant of the estimators' covariance matrix, χ^2_{crit} is the critical value for $\chi^2(v)$ and $\Gamma(x)$ is the gamma function. For WT data the average volume of the hyper-ellipse in case of MLE is 0.0024 while the average hyper-ellipse volume in BI is $3.5162 \cdot 10^{-5}$. In Tet TKO the average volume of the hyper-ellipse for ML estimates is 0.0480 while in case of BI only $9.6 \cdot 10^{-4}$.

k-error clustering

The k-error clustering is a smart modification of the k-means algorithm taking additionally into account the uncertainties of each data point (Kumar and Patel, 2007). Let $\mathbf{v}_1, \ldots, \mathbf{v}_N \in \mathbb{R}^v$ be i.i.d. estimated parameter vectors and $\Sigma_1, \ldots, \Sigma_N \in \mathbb{R}^{v \times v}$ their estimated covariance matrices for all input CpGs. We assume that each parameter vector follows a $v-$ variate normal distribution with one of k possible means $\theta_1, \ldots, \theta_k$, that is $\mathbf{v}_i \sim N_p(\mu_i, \Sigma_i)$, where $\mu_i \in \{\theta_1, \ldots, \theta_k\}$ for $i = 1, \ldots N$. Our goal is to find the clusters $C_1, \ldots, C_k$ such that all the parameter vectors having the same mean $\mu_i = \theta_j$ belong to the same cluster C_j, for $j = 1\ldots,k$.

Let $S_j = \{i | \mathbf{v}_i \in C_j\}$, hence $\mu_i = \theta_j$ for $j = 1\ldots k$ and $\forall i \in S_j$. Given N parameter vectors $\mathbf{v} = (\mathbf{v}_1, \ldots, \mathbf{v}_N)$ and their error matrices $\Sigma_1, \ldots, \Sigma_N$ we search for a partition $S = (S_1, \ldots, S_k)$ and $\theta = (\theta_1, \ldots, \theta_N)$ that maximizes the likelihood: $\mathcal{L}_c(\mathbf{v}) = \prod_{j=1}^k \prod_{i \in S_j} \frac{1}{2\pi}^{p/2} |\Sigma_i|^{-1/2}$ $e^{-1/2(\mathbf{v}_i - \theta_j)\Sigma_i^{-1}(\mathbf{v}_i - \theta_j)^\top}$, where $|\Sigma_i|$ is the determinant of matrix Σ_i for $i = 1, \ldots, N$. It can be shown that maximizing the above likelihood is equivalent as minimizing the total squared Mahalanobis distance of the points that belong to a cluster from the cluster centroid, i.e.,

$$\min_S \sum_{j=1}^k \sum_{i \in S_j} (\mathbf{v}_i - \widehat{\theta}_j) \Sigma_i^{-1} (\mathbf{v}_i - \widehat{\theta}_j),$$

Where $\widehat{\theta}_j$ is the ML estimate of θ_j given by $\widehat{\theta}_j = \left(\sum_{i \in S_j} \Sigma_i^{-1} \right)^{-1} \left(\sum_{i \in S_j} \mathbf{v}_i \Sigma_i^{-1} \right)$ for $j = 1, \ldots, k$. Note that the estimated centroid $\widehat{\theta}_j$ is a weighted mean of the point in cluster C_j, i.e., the Mahalanobis mean of C_j.

After randomly choosing an initial set of k centroids (Forgy method) the k-error algorithm follows as an iteration over the next two steps until no change happens to the assignment of the points.

1. Assign each data point $\boldsymbol{x}_i$ to the cluster whose centroid is the closest using the squared Mahalanobis distance, i.e.,

$$\underset{j}{\operatorname{argmin}} \left(d_{i,j} \right) = \underset{j}{\operatorname{argmin}} (x_i - \widehat{\theta}_j) \Sigma_i^{-1} (x_i - \theta_j)^\top. \qquad \text{(Equation 2)}$$

2. For clusters $\boldsymbol{C}_1, \ldots, \boldsymbol{C}_k$ compute the new cluster centroids $\widehat{\theta}_1, \ldots, \widehat{\theta}_k$ as the Mahalanobis means of the clusters.

To identify the "optimal" number of clusters we use Davies-Bouldin and Calinski-Harabasz criteria that maximize the overall within over between cluster variability. As a distance function we plugin to the above criteria the squared Mahalanobis distance $d(x,y) = (x-y)^\top \Sigma_x^{-1}(x-y)$, where Σ_x is the covariance matrix of point x. From both criteria we obtain two as the optimal number of clusters.

Spatial correlation of enzymatic activity

Let X_s be the discrete space random process describing the dispersion of an enzymatic activity over the whole genome at a certain time point. For a space interval τ its spatial autocorrelation is defined as $R(\tau) = \frac{\mathbb{E}[(X_s - \mu_{X_s})(X_{s+\tau} - \mu_{X_{s+\tau}})]}{\sigma_{X_s}\sigma_{X_{s+\tau}}}$. Similarly the spatial cross-correlation between two random processes X, Y that describe the dispersion of two different enzymatic activities over the genome is defined as $\rho_{X,Y} = \frac{\mathbb{E}[(X_s - \mu_{X_s})(Y_{s+\tau} - \mu_{Y_{s+\tau}})]}{\sigma_{X_s}\sigma_{Y_{s+\tau}}}$. We compute the sample spatial autocorrelation $\widehat{R}$ and the cross-correlations $\widehat{\rho}$ for all enzymatic processes in both WT and Tet TKO experiments as follows. Let genome position $s \in S(\tau)$ when both CpGs of positions s and $s+\tau$ are included in our data. Then

$$\widehat{R}(\tau) = \frac{1}{|S(\tau)-1|\widehat{\sigma}_{X_s}\widehat{\sigma}_{X_{s+\tau}}} \sum_{s \in S(\tau)} (X_s - \overline{X}_s)(X_{s+\tau} - \overline{X}_{s+\tau})].$$

In the above sample estimator $\overline{X_s}$ is the sample mean, and $\widehat{\sigma}_{X_s}$ the sample standard deviation of all measurements X_s for which $s \in S(\tau)$. In the same way we compute

$$\widehat{\rho}(\tau) = \frac{1}{|S(\tau)-1|\widehat{\sigma}_{X_s}\widehat{\sigma}_{Y_{s+\tau}}} \sum_{s \in S(\tau)} (X_s - \overline{X}_s)(Y_{s+\tau} - \overline{Y}_{s+\tau})].$$

Fixing $\tau = 5$ we plot in Figures 3E and S5 the sample autocorrelations and sample cross-correlations between all efficiencies at all time points in WT and Tet TKO experiments. In addition, we report 95% confidence intervals following the approach of (Shen and Lu, 2006) and p-values for the null hypothesis that the auto or the cross-correlation is zero.