A reassessment of DNA-immunoprecipitation-based genomic profiling

A reassessment of DNA immunoprecipitation-based genomic profiling

Antonio Lentini[1], Cathrine Lagerwall[1], Svante Vikingsson[2,3], Heidi K. Mjoseng[4], Karolos Douvlataniotis[1], Hartmut Vogt[1], Henrik Green[2,3], Richard R. Meehan[4], Mikael Benson[1,5] & Colm E. Nestor[1,5]

Affiliations

[1]Division of Pediatrics, Department of Clinical and Experimental Medicine, Linköping University, Linköping, SE58185 Sweden.

[2]Division of Drug Research, Department of Medical and Health Sciences, Linköping University, Linköping, SE58183 Sweden

[3]Department of Forensic Genetics and Forensic Toxicology, National Board of Forensic Medicine, SE58158 Linköping, Sweden.

[4]MRC Human Genetics Unit at the Institute of Genetics and Molecular Medicine at the University of Edinburgh, Crewe Road, Edinburgh, EH4 2XU, UK

[5]These authors contributed equally to this work

Correspondence should be addressed to C.E.N. (colm.nestor@liu.se)

DNA immunoprecipitation sequencing (DIP-seq) is a common enrichment method for profiling DNA modifications in mammalian genomes. However, DIP-seq profiles often exhibit significant variation between independent studies of the same genome and from profiles obtained by alternative methods. Here we show that these differences are primarily due to intrinsic affinity of IgG for short unmodified DNA repeats. This pervasive experimental error accounts for 50 - 99% of regions identified as 'enriched' for DNA modifications in DIP-seq data. Correction of this error profoundly alters DNA modification profiles for numerous cell types, including mouse embryonic stem cells, and subsequently reveals novel associations between DNA modifications, chromatin modifications and biological processes. We conclude that both matched Input and IgG controls are essential to correctly interpret the results of DIP-based assays and that complementary, non-antibody based techniques be used to validate DIP-based findings to avoid further misinterpretation of genome-wide profiling data.

The ability to establish and maintain DNA methylation patterns is essential for normal development in mammals, and aberrant DNA methylation is observed in numerous diseases, including all forms of cancer[1]. Comprehensive mapping of DNA methylation (5-methylcytosine, 5mC) in multiple species has been critical to establishing the relevance of methylation dynamics to gene regulation and chromatin organization[2-4]. An effective method of generating genome-wide 5mC profiles couples antibody-based enrichment of methylated DNA fragments (MeDIP) with hybridization to DNA micro-arrays (MeDIP-chip) or high-throughput sequencing (MeDIP-seq)[5, 6]. MeDIP-seq information is not contained in the read sequence itself, but in the enrichment or depletion of sequencing reads that map to specific regions of the genome[7, 8]. Consequently, appropriate control samples are required, which typically correspond to the input genomic DNA before enrichment. More recently, DIP-seq has been extended to chart the genomic location of additional DNA modifications including 5-hydroxymethylcytosine (5hmC), 5-formylcytosine (5fC), 5-carboxycytosine (5caC) and 6-

methyladenosine (6mA). Verification of DIP profiles by independent methods revealed several problems with the DIP-seq approach, including preferential enrichment of low CG content regions by the 5mC antibody[9] and enrichment of highly modified regions by the 5hmC antibody[10]. In addition, we and others have reported high background signals in 5hmC DIP assays[11-14] which was partly due to non-specific enrichment of short tandem repeats (STRs)[11, 12]. However, the origin of STR enrichment and the scale of its impact on DIP-seq data remained unknown.

Here, we demonstrate that highly specific off-target binding to unmodified STRs is not limited to 5hmC antibodies but is an inherent technical error observed in all DIP-seq studies, irrespective of the target DNA modification, cell-type or organism. We reveal that between 50% - 99% of enriched regions in DIP-Seq data are false positives, the removal of which markedly affects our perception of methylation dynamics in mammals. Our findings will substantially improve the accuracy of future DIP-seq experiments and allow new insights to be gained from the wealth of existing DIP-seq data.

RESULTS

IgG antibodies have an intrinsic affinity for short tandem repeats in mammalian DNA

To simplify comparison of DIP-seq results from separate studies we used a uniform computational pipeline (see **online methods**) to analyze published DIP-seq profiles of 5mC, 5hmC, 5fC and 5caC (hereby referred to as '5modC') in mouse embryonic stem cells (mESCs). All analyzed datasets and their relationship to figures is outlined in **Supplementary Table 1**. This approach revealed a striking enrichment at short tandem repeats (STRs) in all 5modC DIP-seq datasets (**Fig. 1a** and **Supplementary Fig. 1**). This could not be explained by non-specific binding of the antibodies to other modifications as the specificity of antibodies used in DIP-seq is well established[11, 12, 15] and was confirmed by dot-blot and ELISA assays for commercially available antibodies (**Supplementary Fig. 2a, b**). Surprisingly, near identical enrichment patterns at STRs were observed in mESC DIP-seq generated with a non-specific mouse IgG antibody (**Fig. 1a** and **Supplementary Fig. 1**). The intersection of regions enriched for all 5modC showed a 19 fold higher enrichment for IgG compared to Input (median RPM = 0.824 and 0.043 for IgG and Input, respectively; $P=5.03\text{x}10^{-5}$, T-test) whereas non-intersecting regions showed no difference (**Supplementary Fig. 2c**), suggesting that a proportion of the 5modC signal may be due to off-target binding of the antibodies. Indeed, genome-wide IgG enrichment could explain up to 55% of all 5modC DIP-seq enriched loci in mESCs whereas Input explained a maximum of 3% of enriched regions (**Supplementary Fig. 2d**). Significantly, overlapping 5mC, 5hmC and IgG regions were depleted of CpG dinucleotides compared to regions not overlapping IgG (**Supplementary Fig. 2e**). Although non-CpG methylation is known to occur in mESCs[16, 17], analysis of whole-genome bisulfite sequencing data[16] confirmed that CpHs in these regions were primarily unmethylated (median methylated CpHs = 0 and 8 for IgG and 5mC regions, respectively; $P<1\text{x}10^{-16}$, Mann-Whitney U-test) (**Supplementary Fig. 2f**) suggesting that all antibodies were non-specifically binding regions of unmodified DNA during DIP experiments. We verified this by analyzing published DIP-seq data from DNMT triple knockout (TKO) mESCs[18] that lack DNA methyltransferase activity and revealed that both the 5mC and 5hmC antibodies enriched similar regions to that of the IgG control in these samples (**Fig. 1b** and **Supplementary Fig. 2g**). This was further reinforced by 5hmC DIP-seq profiles from mouse embryoid bodies lacking all three *TET* genes with

undetectable levels of 5hmC[19] (**Supplementary Fig. 2g**). We confirmed depletion of both 5mC and 5hmC in DNMT TKO compared to wild-type (WT) mESC DNA using mass spectrometry (**Fig. 1c**), verifying that the DIP-seq signals observed in TKO cells were independent of 5modC status. 5hmC-DIP followed by qPCR confirmed the enrichment of STRs in TKO mESCs lacking 5hmC (**Fig. 1d**). Significantly, 5hmC profiles generated from an independent, non-antibody based 5hmC enrichment technique[20] (5hmC-Seal) showed no enrichment over IgG regions (**Fig. 1e**) further implicating off-target binding of STRs by antibodies during DIP-seq. Importantly, the observation that 5hmC-Seal does not enrich for STRs despite using an identical PCR amplification protocol to that of 5hmC-DIP, excludes PCR amplification as the source of the observed STR enrichment (**Fig. 1e** and **Supplementary Fig. 2h,i**)[15, 20].

To identify specific IgG-bound sequences, we screened the raw sequencing reads from three IgG DIP-seq samples in mESCs for overrepresented sequences, which revealed that between 30 and 60% of all reads were significantly enriched for repetitive motifs compared to Input (**Fig. 1f** and **Supplementary Table 2**), including the previously reported CA-repeats[11]. This suggested that IgG antibodies may have an innate binding capacity for repetitive DNA sequences. Not only were IgG DIP-seq enriched for repetitive motifs, but the enriched IgG motifs were highly similar between samples (average Pearson r = 0.72) indicating that IgG binding is specific and reproducible (**Supplementary Table 2**). We observed similar repeat motifs in 5modC DIP-seq data from mESCs as well as a recently published study in mouse embryonic fibroblasts (MEFs)[21] (r mESC = 0.75, r MEF = 0.68, **Supplementary Table 2**), showing that off-target binding of STRs in DIP-seq is not limited to mESCs and is highly sequence dependent. Indeed, the only antibody-based profiling technique that did not show enrichment over IgG enriched regions was cytosine-5-methylenesulfonate (CMS)-seq[22] (**Fig. 1e**), which involves bisulfite conversion of all unmodified cytosines to thymine before immunoprecipitation with the anti-CMS antibody. Consequently, all unmodified CA-repeats would be converted to TA-repeats. The lack of IgG enrichment in anti-CMS is thus strongly supportive of sequence-specific off-target binding of STRs by IgG antibodies. Taken together, our analyses indicates that native DNA immunoprecipitation libraries generated with multiple cytosine modification antibodies enriches for highly specific sequences of unmodified repetitive DNA.

IgG binding of DNA repeats and bacterial contamination explains the conflicting results of 6mA profiling in vertebrates

Next, we extended our analysis to a non-cytosine modification, 6-methyldeoxyadenosine (6mA), that is abundant in many bacteria and recently characterized in invertebrates[23-27]. Its subsequent discovery in mammalian DNA has sparked an intense research effort to verify its location and characterize its function[27-30] however the existence of 6mA in mammals remains controversial[31-33]. To determine if 6mA DIP-seq studies have also been affected by off-target IgG binding we compared 6mA DIP-seq profiles from mESCs[28], primary mouse kidney cells[27] and mouse prefrontal cortex (hereby referred to as 'brain')[30] to mESC IgG DIP-seq profiles. Again, 6mA profiles showed a clear enrichment at STRs and IgG enriched regions in mESCs (**Fig. 2a,b**). We next compared enriched 6mA regions with data from DIP-seq in DNMT TKO cells and found that not only was the enrichment for STRs highly similar but it also differed significantly from both Input and 5hmC (P=0.54, $1.6\text{x}10^{-3}$ and $1.1\text{x}10^{-5}$ for TKO, Input and 5hmC, respectively, 'BH' corrected T-tests) (**Fig. 2c**). This means that DIP-seq using a specific 6mA antibody in mice resulted in near identical enrichment as using random antibodies in tissues lacking the target modifications, suggesting that the 6mA DIP-seq signal in mice is

mainly mediated by off-target IgG binding. Analysis of additional public datasets in multiple species revealed that 6mA DIP-seq data for *Danio rerio*[29] and *Xenopus laevis*[27] also showed similar off-target enrichment for the same STR motifs observed in 5modC DIP-seq, albeit at a lower degree, whereas the 6mA rich genomes of *C. elegans*[25] and *E. coli*[27] showed no enrichment for these motifs (**Fig. 2d, e** and **Supplementary Table 3**). Correlation with IgG motifs in mESCs reflected the inter-species frequency of CA-repeats in the different genomes (**Fig. 2f, Supplementary Fig. 3a**), showing that off-target binding will vary greatly between species due to inter-species differences in STR composition. We next identified 6mA enriched regions in *X. laevis* genome-wide for three different antibodies (N=2) using Input controls, yielding on average 24,540 enriched regions which was highly similar to what was reported in the original publication[27]. However, when controlling for IgG, the number of identified enriched regions was reduced to 256 on average, meaning that > 98% of all Input-identified regions were not detectable when using IgG controls (**Supplementary Fig. 3b**). This implies that nearly all of the 6mA signal was due to off-target binding of IgG. Furthermore, caution has been raised regarding cell culture contamination[32, 33] as common bacterial contaminants contain high levels of 6mA and other DNA modifications[34, 35]. To test this we classified sequencing reads to a combined genome index of *M. musculus* and common cell culture contaminants (**see Online Methods**). This revealed substantial contamination of several DIP-seq datasets with bacterial DNA including *Mycoplasma spp.*. Notably, the proportion of bacterial read contamination differed substantially between 6mA DIP-Seq of WT and ALKBH1 KO mESCs[28] (**Supplementary Fig. 3c** and **Supplementary Table 4**). We further tested 21 different 5modC DIP-seq samples used throughout our analysis which showed no evidence for *Mycoplasma spp.* contamination (**Supplementary Table 4**). Contamination of these samples may explain the earlier detection of 6mA in mESCs by mass spectrometry[28] and the subsequent failure of more recent attempts using ultrasensitive UHPLC-MS[31].

Normalizing for off-target IgG binding sharpens our view of epigenetic organization in mammals

To determine how off-target binding in DIP-seq has affected our understanding of DNA methylation in mammals, we reanalyzed data from five independent studies of 5modC marks in mESCs[11, 15, 18, 36, 37]. First, we determined the fraction of false positive regions when using Input as a control (**Supplementary Fig. 4a**), finding that up to 99% of enriched 5fC and 5caC, and approximately half of all 5hmC and 5mC regions could be considered false positives (**Fig. 3a**). In contrast, the mean percentage of falsely enriched regions was approximately 7% on average for all 5modC marks when using IgG as a control (**Fig. 3a**). Since suppression of *Tdg* markedly increases levels of 5caC and 5fC[15, 21], we also determined the false positive rate for *Tdg* knockdown in mESCs and found that whereas falsely enriched regions using IgG remained constant around 5% on average, using Input controls decreased false positive rates by around 50% and 25% for 5caC and 5fC, respectively, while 5hmC and 5mC remained largely unchanged (**Supplementary Fig. 4b**) clearly showing that off-target binding is relative to mark abundance. These results suggested that not only is Input a highly inconsistent control but also that the 5modC landscape in mammalian genomes has been greatly overestimated by DIP-seq (**Supplementary Fig. 4c**). Indeed, correcting for IgG not only reduced the number of enriched regions but also greatly increased the overlap with anti-CMS and Seal profiling techniques (**Supplementary Fig. 4d**). Not surprisingly, the proportion of enriched repeat types was markedly altered when using Input or IgG controls in DIP-Seq, with STRs showing changes in enrichment for all marks but 5fC (**Supplementary Fig. 4e**). Interestingly, whereas enrichment in AG-repeats was lower for all marks, over 30% of all 5fC enriched regions were in CA-repeats even after correcting for IgG (**Supplementary Fig. 4f**) suggesting biological importance of 5fC at CA-repeats. Indeed, a recent study showed that 5fC at intronic CA-repeats

was associated with gene silencing[21] underlining the biological importance of modifications of repetitive elements in gene regulation.

Globally, 49% of 5mC- co-located with 5hmC enriched regions when using Input, whereas only 17% were coincident for both 5hmC and 5mC when using IgG (**Fig. 3b**). This suggested a more restricted role for 5hmC mediated DNA de-methylation in the reprogramming of the mESC epigenome, an assertion supported by the markedly improved association between 5hmC and TET protein occupancy in the mESC genome upon normalization to IgG (**Fig. 3c**). Significantly, removal of signals caused by off-target binding by normalization to IgG also altered the association of 5hmC with biological pathways from non-significant associations with unrelated processes including 'cilia formation', 'smell perception' and 'phosphorus metabolism' to highly significant associations with processes related to mammalian development and cell differentiation (**Fig. 3d, upper panels**). Significantly, the 5hmC-associated biological processes identified after correction for STR-binding were highly similar to those obtained with 5hmC-Seal and anti-CMS, which do not enrich for unmodified repeats (**Fig. 3d, lower panels**). An improved association with developmental and differentiation related processes was also observed when the same correction was applied to MEFs (**Supplementary Fig. 4g**).

Finally, histone ChIP-seq data in mESCs from ENCODE[38] showed no enrichment over IgG DIP-seq enriched regions (**Fig. 3e** and **Supplementary Fig. 4h**) suggesting that repeats found in intact chromatin structures are not bound by IgG, possibly due to their inability to form secondary structures. Again, using an IgG control significantly increased the association of 5hmC with permissive histone marks in mESCs[38] whereas the association with heterochromatin (H3K9me3) decreased (**Fig. 3f**). For 5mC, the association with histone marks was also significantly increased, accentuating co-localization with heterochromatin (H3K9me3) as well as H3K36me3 which together with 5mC is involved in mRNA splicing[39] (**Fig. 3f**).

DISCUSSION

Our reanalysis of published DIP-seq data revealed that all commonly used DIP-seq antibodies bind unmodified short tandem repeat (STR) sequences. By analyzing DIP-seq data from mouse embryonic stem cells (mESCs) lacking both 5mC and 5hmC we confirmed that STR binding was modification-independent. Consequently, only studies that have normalized DNA modification enrichment to an IgG control have corrected for off-target binding[15, 23] (**Fig. 4**). Unfortunately, 95% of published DIP-seq studies (unique DIP-Seq studies in the GEO database, January 2018) do not include an IgG control. We show that between 50 to 99% of enriched regions are due to off-target binding in 5modC DIP studies. Off-target binding was highly related to abundance of the target with low abundance modifications (i.e. 5caC & 5fC) having the highest false positive rates which could be effectively altered by increasing 5caC and 5fC levels through TDG knockdown. This means that not only does Input not control for off-target binding but is also highly inconsistent between DIP experiments of different targets, species, and tissues. Controlling for off-target IgG binding increased the signal-to-noise ratio in DIP-seq assays >3-fold, allowing identification of more subtle alterations in modification levels. This also results in a significantly smaller and more distinct epigenomic landscape in mammalian cells, evidenced by a significantly reduced overlap between 5mC and 5hmC marked loci and a stronger association between 5modC and a variety of chromatin marks. Thus, IgG DIP-seq controls and validation of enrichment by independent (non-DIP) techniques are

essential for appropriate interpretation of future DIP-seq experiments (see **Supplementary Discussion**)

Unexpectedly, we also revealed the potential for contaminating bacterial DNA to confound the results of DIP-seq studies of trace DNA modifications. The risk of such contaminants has been previously raised with regards to 6mA[23, 40], which is vanishingly rare in mammals, but highly abundant in many bacterial species that commonly infect mammalian cell cultures, such as *Mycoplasma* and *E.coli*. Fortunately, even minor bacterial contamination of mammalian DNA samples can be identified by comparison of next generation sequencing reads with the genomic sequence of suspected contaminants. Using this approach, we found that up to 17% of reads in published samples of DIP-seq datasets in mammals mapped to the *Mycoplasma* genome. Moreover, the proportion of bacterial read contamination often differed substantially between DIP-seq datasets of test samples and their matched control samples, severely undermining observations of altered 6mA content and distribution between experimental conditions[28]. Taken together with the results of a recent study that was unable to detect 6mA in mammalian cells using mass spectrometry[31] and our results showing clear IgG off-target binding using the 6mA antibody, a re-evaluation of the extent and origin of 6mA in mammalian studies is advisable.

How non-specific DNA molecules become bound to IgG during DNA immunoprecipitation is unclear. Interestingly, whereas the 5mC enrichment-based MethylCap technique utilizing a MBD-GST fusion protein does not show enrichment for STRs[11], the use of a MBD-Fc fusion protein shows specific enrichment of both CA- and AG-repeats[41] suggesting that off-target binding of repeats is mediated by the Fc region of IgG. As DNA is typically denatured prior to immunoprecipitation, it is tempting to speculate that ssDNA molecules may bind directly to the conserved Fc region of IgG antibodies. Indeed, both ssRNA and ssDNA molecules ('aptamers') capable of specifically binding the Fc-region of mouse and rabbit IgG have been reported[42]. However, although DNA is denatured prior to immunoprecipitation, high copy number repeats rapidly re-associate during the cooling process[43]. Thus, the denatured DNA samples used in DIP are likely to contain a significant proportion of double stranded repetitive sequences, making it difficult to conclude from the current data whether IgG binding of STRs is sequence or structure dependent. Regardless of the mechanistic underpinnings of STR enrichment during DIP, a matched IgG control will normalize for off-target binding in all cases. Whereas our discovery of unmodified STR binding by IgG has revealed a serious flaw in DIP-seq to date, it will allow the field to minimize the impact of these errors on future DIP based assays and accelerate the discovery of novel findings from the multitude of existing DIP-seq data.

ACCESSION CODES

The sequencing data analyzed in this study are publicly available through GEO or ENA under accessions GSE4225062, GSE2484363, GSE3134364, ERP00057065, GSE2850066, GSE7186667, GSE7418468, GSE7674069, GSE7954370, GSE6650471, GSE5504972, GSE4192373, GSE4154574, GSE2868275 and mouse ENCODE49 data is available from https://www.encodeproject.org/.

DATA AVAILABILITY

The sequencing data analyzed in this study are publicly available through GEO or ENA under accessions GSE4225062, GSE2484363, GSE3134364, ERP00057065, GSE2850066, GSE7186667, GSE7418468, GSE7674069, GSE7954370, GSE6650471, GSE5504972, GSE4192373, GSE4154574, GSE2868275 and mouse ENCODE49 data is available from https://www.encodeproject.org/.

See **Supplementary Table 1** for specification of files used for each analysis/figure.

ACKNOWLEDGEMENTS

Work in the lab of C.E.N was supported by the Swedish Research Council (2015-03495), LiU-Cancer (2016-007) and the Swedish Cancer Society (CAN 2017/625). R.R.M. and H.K.M. were supported by the Medical Research Council, UK (MC_PC_U127574433). M.B. was supported by the Swedish Research Council (2015-02575). H.G. was supported by the Swedish Cancer Society (CAN 2016/602).

AUTHOR CONTRIBUTIONS

C.L., S.V., K.D., and H.K.M. performed experiments, A.L., C.E.N. and S.V. analyzed data, A.L. R.R.M. and C.E.N. wrote the manuscript and H.V., H.G., R.R.M., M.B. and C.E.N. supervised the work.

COMPETING FINANCIAL INTERESTS

The authors declare no conflicts of interest.

REFERENCES

1. Goll, M.G. & Bestor, T.H. Eukaryotic cytosine methyltransferases. *Annual review of biochemistry* **74**, 481-514 (2005).
2. Bogdanovic, O. et al. Active DNA demethylation at enhancers during the vertebrate phylotypic period. *Nature genetics* **48**, 417-426 (2016).
3. Feinberg, A.P. & Tycko, B. The history of cancer epigenetics. *Nat Rev Cancer* **4**, 143-153 (2004).
4. Illingworth, R.S. et al. Orphan CpG islands identify numerous conserved promoters in the mammalian genome. *PLoS genetics* **6**, e1001134 (2010).
5. Weber, M. et al. Chromosome-wide and promoter-specific analyses identify sites of differential DNA methylation in normal and transformed human cells. *Nature genetics* **37**, 853-862 (2005).
6. Harris, R.A. et al. Comparison of sequencing-based methods to profile DNA methylation and identification of monoallelic epigenetic modifications. *Nature biotechnology* **28**, 1097-1105 (2010).
7. Bock, C. Analysing and interpreting DNA methylation data. *Nature reviews. Genetics* **13**, 705-719 (2012).
8. Bock, C. et al. Quantitative comparison of genome-wide DNA methylation mapping technologies. *Nature biotechnology* **28**, 1106-1114 (2010).
9. Nair, S.S. et al. Comparison of methyl-DNA immunoprecipitation (MeDIP) and methyl-CpG binding domain (MBD) protein capture for genome-wide DNA methylation analysis reveal CpG sequence coverage bias. *Epigenetics* **6**, 34-44 (2011).
10. Ko, M. et al. Impaired hydroxylation of 5-methylcytosine in myeloid cancers with mutant TET2. *Nature* **468**, 839-843 (2010).

11. Matarese, F., Carrillo-de Santa Pau, E. & Stunnenberg, H.G. 5-Hydroxymethylcytosine: a new kid on the epigenetic block? *Molecular systems biology* **7**, 562 (2011).
12. Thomson, J.P. et al. Comparative analysis of affinity-based 5-hydroxymethylation enrichment techniques. *Nucleic acids research* **41**, e206 (2013).
13. Skvortsova, K. et al. Comprehensive evaluation of genome-wide 5-hydroxymethylcytosine profiling approaches in human DNA. *Epigenetics & chromatin* **10**, 16 (2017).
14. Pastor, W.A., Huang, Y., Henderson, H.R., Agarwal, S. & Rao, A. The GLIB technique for genome-wide mapping of 5-hydroxymethylcytosine. *Nature protocols* **7**, 1909-1917 (2012).
15. Shen, L. et al. Genome-wide analysis reveals TET- and TDG-dependent 5-methylcytosine oxidation dynamics. *Cell* **153**, 692-706 (2013).
16. Habibi, E. et al. Whole-genome bisulfite sequencing of two distinct interconvertible DNA methylomes of mouse embryonic stem cells. *Cell stem cell* **13**, 360-369 (2013).
17. Ramsahoye, B.H. et al. Non-CpG methylation is prevalent in embryonic stem cells and may be mediated by DNA methyltransferase 3a. *Proceedings of the National Academy of Sciences of the United States of America* **97**, 5237-5242 (2000).
18. Williams, K. et al. TET1 and hydroxymethylcytosine in transcription and DNA methylation fidelity. *Nature* **473**, 343-348 (2011).
19. Dawlaty, M.M. et al. Loss of Tet enzymes compromises proper differentiation of embryonic stem cells. *Developmental cell* **29**, 102-111 (2014).
20. Song, C.X. et al. Genome-wide profiling of 5-formylcytosine reveals its roles in epigenetic priming. *Cell* **153**, 678-691 (2013).
21. Papin, C. et al. Combinatorial DNA methylation codes at repetitive elements. *Genome research* **27**, 934-946 (2017).
22. Pastor, W.A. et al. Genome-wide mapping of 5-hydroxymethylcytosine in embryonic stem cells. *Nature* **473**, 394-397 (2011).
23. Traube, F.R. & Carell, T. The chemistries and consequences of DNA and RNA methylation and demethylation. *RNA Biol*, 1-9 (2017).
24. Fu, Y. et al. N6-methyldeoxyadenosine marks active transcription start sites in Chlamydomonas. *Cell* **161**, 879-892 (2015).
25. Greer, E.L. et al. DNA Methylation on N6-Adenine in C. elegans. *Cell* **161**, 868-878 (2015).
26. Zhang, G. et al. N6-methyladenine DNA modification in Drosophila. *Cell* **161**, 893-906 (2015).
27. Koziol, M.J. et al. Identification of methylated deoxyadenosines in vertebrates reveals diversity in DNA modifications. *Nature structural & molecular biology* **23**, 24-30 (2016).
28. Wu, T.P. et al. DNA methylation on N(6)-adenine in mammalian embryonic stem cells. *Nature* **532**, 329-333 (2016).
29. Liu, J. et al. Abundant DNA 6mA methylation during early embryogenesis of zebrafish and pig. *Nat Commun* **7**, 13052 (2016).
30. Yao, B. et al. DNA N6-methyladenine is dynamically regulated in the mouse brain following environmental stress. *Nat Commun* **8**, 1122 (2017).
31. Schiffers, S. et al. Quantitative LC-MS Provides No Evidence for m6 dA or m4 dC in the Genome of Mouse Embryonic Stem Cells and Tissues. *Angewandte Chemie* (2017).
32. Luo, G.Z. & He, C. DNA N(6)-methyladenine in metazoans: functional epigenetic mark or bystander? *Nature structural & molecular biology* **24**, 503-506 (2017).

33. O'Brown, Z.K. & Greer, E.L. N6-Methyladenine: A Conserved and Dynamic DNA Mark. *Advances in experimental medicine and biology* **945**, 213-246 (2016).
34. Razin, A. & Razin, S. Methylated bases in mycoplasmal DNA. *Nucleic acids research* **8**, 1383-1390 (1980).
35. Lluch-Senar, M. et al. Comprehensive methylome characterization of Mycoplasma genitalium and Mycoplasma pneumoniae at single-base resolution. *PLoS genetics* **9**, e1003191 (2013).
36. Ficz, G. et al. Dynamic regulation of 5-hydroxymethylcytosine in mouse ES cells and during differentiation. *Nature* **473**, 398-402 (2011).
37. Xu, Y. et al. Genome-wide regulation of 5hmC, 5mC, and gene expression by Tet1 hydroxylase in mouse embryonic stem cells. *Molecular cell* **42**, 451-464 (2011).
38. Yue, F. et al. A comparative encyclopedia of DNA elements in the mouse genome. *Nature* **515**, 355-364 (2014).
39. Brown, S.J., Stoilov, P. & Xing, Y. Chromatin and epigenetic regulation of pre-mRNA processing. *Human molecular genetics* **21**, R90-96 (2012).
40. Luo, G.Z., Blanco, M.A., Greer, E.L., He, C. & Shi, Y. DNA N(6)-methyladenine: a new epigenetic mark in eukaryotes? *Nature reviews. Molecular cell biology* **16**, 705-710 (2015).
41. Gebhard, C. et al. General transcription factor binding at CpG islands in normal cells correlates with resistance to de novo DNA methylation in cancer cells. *Cancer research* **70**, 1398-1407 (2010).
42. Nezlin, R. Aptamers in immunological research. *Immunology letters* **162**, 252-255 (2014).
43. Waring, M. & Britten, R.J. Nucleotide Sequence Repetition: A Rapidly Reassociating Fraction of Mouse DNA. *Science* **154**, 791-794 (1966).

FIGURE LEGENDS

Figure 1. Characterization of off-target antibody binding in DIP-seq. (**a**) Signal track in mESCs showing similar enrichment between 5modC and IgG DIP-seq samples over repetitive regions. WGBS, whole-genome bisulfite sequencing; STRs, short tandem repeats. (**b**) Signal track of 5mC, 5hmC and IgG DIP-seq in DNMT triple knockout (TKO) or wild-type (WT) mESCs over 5hmC- (left) or IgG enriched regions (right). (**c**) Mass spectrometry quantification of 5mC and 5hmC in TKO and WT mECSs for $n = 3$ biologically independent samples. Data shown as mean ±s.d. P-values calculated using two-tailed T-test. (**d**) DIP using a 5hmC antibody in wild-type (WT) (left) and $DNMT^{TKO}$ (right) mESCs for DIP-qPCR $n = 3$ and DIP-seq $n = 1$ biologically independent samples. Data represented as in **c**. Correlation between mean DIP-qPCR and DIP-seq values calculated using two-tailed Spearman correlation. STRs, short

tandem repeats. (**e**) 5hmC enrichment in mESCs with different profiling techniques over 5hmC n = 31265 enriched regions (left) or IgG n = 137557 enriched regions (right). (**f**) Consensus motif enrichment for raw IgG reads compared to Input of n = 3 biologically independent samples.

Figure 2. Characterization of similarities between 6mA and IgG DIP-seq in different species. (**a**) Signal track for Input and 6mA DIP-seq in mouse tissues and IgG DIP-seq in mESCs. STRs, short tandem repeats. (**b**) Enrichment over IgG enriched DIP-seq regions for 6mA DIP-seq n = 11 and Input n = 4 biologically independent samples. P-values calculated using two-tailed T-test. Boxplots represent median and first and third quartiles with whiskers extending 1.5 * inter-quartile range. (**c**) Fraction of DIP-seq enriched regions located in short tandem repeats (STRs) for 6mA n = 11, TKO n = 3, Input n = 4 and 5hmC n = 6 biologically independent samples. P-values calculated from biologically independent samples using pairwise two-tailed T-tests with Benjamini-Hochberg correction for multiple testing. Data represented as in **b**. (**d**) Motif enrichment for raw 6mA or IgG DIP-seq reads compared to Input in multiple species. Motif with highest correlation to IgG motifs shown for each cell type and antibody. (**e**) Fraction of motifs highly similar (r > 0.75) to mouse IgG motifs for *M. musculus* n = 11, *D. rerio* n = 2, *X. laevis* n = 8, *C. elegans* n = 1 and *E. coli* n = 2 biologically independent samples. Data represented as in **b**. (**f**) Proportion of CA-repeats in the genomes of model organisms.

Figure 3. Biological impact of IgG correction. (**a**) Estimated false positive rate of DIP-seq enriched regions using IgG or Input as control in mESCs for 5caC n = 2, 5fC n = 2, 5hmC n = 7 and 5mC n = 6 biologically independent samples. Data shown as mean ±s.d. (**b**) Overlap of 5hmC and 5mC regions using IgG or Input controls showing decreased overlap when using IgG controls. Venn diagram of 5mC and 5hmC overlap using IgG or Input controls (top) and paired line plot of 5mC and 5hmC overlap using IgG or Input controls for multiple studies (indicated by symbols, bottom). Data shown as mean and individual data points of n = 6

biologically independent samples. P-values calculated using two-tailed paired T-test. ▲ = ERP000570, ● = GSE31343, ■ = GSE24841, ▼= GSE42250. (**c**) TET1 binding over IgG *n* = 137557 enriched regions or 5hmC *n* = 31265 enriched regions using IgG or Input controls. (**d**) GO term enrichment for top genes (*n* = 500) enriched for 5hmC using DIP-seq with either IgG or Input controls or 5hmC-Seal or anti-CMS techniques. P-values calculated using PANTHER biological processes. (**e**) Relative enrichment of ENCODE mESC histone ChIP-seq data for *n* = 26 biologically independent samples in regions enriched for IgG in DIP-seq or random regions of same size and chromosome. Boxplots represent median and first and third quartiles with whiskers extending 1.5 * inter-quartile range. P-values calculated using two-tailed T-test. (**f**) Enrichment of ENCODE mESC histone ChIP-seq data for 5hmC- (left) or 5mC (right) enriched regions using IgG or Input as controls. Data presented as mean (IgG) and bootstrapped mean (Input) of H3K27ac *n* = 2, H3K36me3 *n* = 4, H3K4me1 *n* = 6, H3K4me3 *n* = 4, H3k9ac *n* = 2, H3K27me3 *n* = 2, H39me3 *n* = 4 biologically independent samples, #P<1e-5, bootstrap resampling (*n* = 100,000).

Figure 4. Antibodies in DIP-seq experiments bind repetitive elements which are incorrectly identified as enriched regions when not controlled for IgG binding.

ONLINE METHODS

Cell culture. J1 mouse embryonic stem cells (mESCs; WT, male) were originally derived from the 129S4/SvJae strain. TKO (Dnmt1-/-, Dnmt3a-/-, Dnmt3b-/-) mESCs were derived from J1 mESCs[44]. Both cell lines were cultured in a humidified incubator at 5% CO_2, 37°C on 0.2% gelatin coated tissue culture plastic in DMEM (Dulbecco's modified eagle medium) supplemented with 15 % fetal calf serum, 0.1 mM non-essential amino acids (Sigma-Aldrich, MI, USA), 1 mM sodium Pyruvate (Sigma-Aldrich, MI, USA), 1 % Penicillin/Streptomycin,

2 mM L-glutamine, 0.1 mM beta-mercaptoethanol (Thermo Fisher, CA, USA), and ESGRO LIF (Millipore, MA, USA) at 500U/mL. mESCs were passaged every 2-3 days using trypsin/EDTA.

DNA extraction. Snap frozen cell pellets were treated with RNAse cocktail (Ambion, CA, USA) for 1 hour at 37°C followed by proteinase K treatment overnight at 55°C. DNA was extracted by standard phenol chloroform/ethanol precipitation and eluted in TE.

DIP-qPCR. 1.5 µg genomic DNA was sonicated to fragments ranging between 100-1000 bp with a peak at 400 bp using a BioRuptor (Diagenode, Belgium), denatured at 95°C for 10 min then cooled on wet ice for 10 min. 10% of samples were saved as Input and the remaining DNA was resuspended in 10x IP buffer (10 mM Na-Phosphate (mono-dibasic), 1% NaCl, 0.05% Triton X-100, pH 7.0). Immunoprecipitations were performed using 1µg anti-5hmC antibody (Active Motif, #39769) for 12h at 4°C using constant rotation. Protein G dynabeads (Invitrogen, CA, USA, #100-03D) were washed twice in 0.1% PBS-BSA then added to the IP mixture for 1h at 4° using constant rotation. Beads were washed three times for 10 min using cold 1x IP buffer then resuspended in digestion buffer and incubated with 8 U Proteinase K (New England Biolabs, MA, USA) for 1.5h at 50°C, 800rpm in 50 mM Tris, 10 mM EDTA 0.5% SDS, pH 8.0 and purified using DNA Clean & Concentrator kit (Zymo Research, USA). Quantitative PCR was performed on a 7900HT real-time cycler (Applied Biosystems, CA, USA) using SYBR green master mix (Applied Biosystems, CA, USA). qPCR primers use are listed in **Supplementary Table 4**, below.

Supplementary Table 4. hMeDIP qPCR primer sequences

name	forward primer (5' - 3')	reverse primer (5' – 3')	designation
Rho	ACCGTACAGCACAAGAAGCTGC	GAAGACCATGAAGAGGTCAGCC	True Positive
Aqp2	ATGTGGGAACTCCGGTCCATAG	GCCAAAGAAGACGAAAAGGAGC	True Positive
ActB	ATGAAGAGTTTTGGCGATGG	GATGCTGACCCTCATCCACT	True Negative
Baiap2l1	ATCTGCACTTGATGACAACTGG	CTTGTGAGACCAAGCTCTTAGC	True Negative
Cyp3a41a	TTCACCTTTATGACTTGGTAGGC	GCTTCTCTTGTGAGGACTGTGG	False Positive
Arpc1a	TGGGGCTCATTTCTGTAATACC	TTCCATCTTCTCAAATCATTGC	False Positive
Nptx2	TCTCAAGTGCTGGGATTAAAGG	TCTGGGAAGCAAATCTAAGTCC	False Positive
Gm4871	CTGGTGTGTGTTTATCCTCAGC	AACTGTGGAGTGAGGTATGAAGG	False Positive
Bri3	TGGAGAGTGTGTATGTGTGAGC	AGGAGGCAGAAGGAGAAAGG	False Positive
Clec4e	CACATACTGCCTTCTGCTATGC	TGTGTGAGTGAAAGGAGAGAGC	False Positive
Kpna7	CAACCAGGACTACACAGTGACG	GACACAGAAGCACAGAGAGAGG	False Positive
Eif2ak	AGAGGCCAGAAGGTGTTGG	TTTCAGAGGACCTGAGTTTGG	False Positive

Quantification of cytosine modifications using mass spectrometry. 1 µg of DNA was heat denatured at 100 °C for 5 min in 20µL H_2O then immediately cooled on ice. 10 µl P1 Nuclease (0.02 U/µl in 90 mM AmAc, 0.3 mM $ZnSO_4$, pH 5.3) was added followed by incubation at 50 °C for 2 h. 10 µl Alkaline phosphatase (0.08 U/µl in 200 mM TRIS-HCl, 0.40 mM EDTA, pH 8) was added followed by incubation at 37 °C for 30 min. Proteins were precipitated by the addition of 160 µl cold acetonitrile. Following centrifugation at 17000 x g for 5 min, 180 µl of the supernatant was evaporated under nitrogen and reconstituted in 40 µl 0.1% formic acid. The chromatographic system consisted of an Acquity UPLC (Waters, MA, USA) and a Xevo triple quadrupole mass spectrometer (Waters, MA, USA). The extracts were separated on an HSS T3 column (150x2.1 mm, 1.7 µm, Waters, MA, USA) at 45°C and a flow rate of 450 µl/min using a gradient elution with 0.05% acetic acid and methanol, 0-1.3 min 2% B; 1.3-5.5 min 2-9% B; 5.5-7.5 min re-equilibration at 2% B. For dC a 1 µl injection was made and for mC, hmC, fC and caC a 15 µl injection was made. Analytes were detected in the multi reaction monitoring (MRM) mode using three time windows with the following transistions 0-2.3 min

– C (228->95 & 228->112) and hmC (258->124 & 258->142); 2.3-4 min – mC (242->109, 242->126) and caC(272->138, 272->156); 4-7.5 min – fC (256->97, 256->140).

Immuno dot-blot. 10 ng 426 bp oligos containing 5mC, 5hmC, 5fC, 5caC or C (GeneTex, CA, USA) was denatured at 95°C for 15 min in 0.4M NaOH and 10mM EDTA then immediately cooled on ice. Samples were applied to a positively charged nylon membrane under vacuum using a Dot Blot Hybridisation Manifold (Harvard Apparatus, MA, USA). The membranes were briefly washed in 2X SSC buffer (0.3M NaCl, 30mM NaCitrate) then cross-linked using a UV Stratalinker 1800 (Stratagene, CA, USA) and baked at 80°C for 2 h. Membranes were blocked in casein blocking buffer (Li-Cor) for 15 min at 4°C then incubated with an antibody against 5mC (1:3000, Zymo #A3001), 5hmC (1:3000, ActiveMotif #39791), 5fC (1:3000, ActiveMotif #61227) or 5caC (1:3000, ActiveMotif #61229) for 1h at 4°C. Membranes were washed 3 times for 5 min in TBS-Tween (0.05%) then incubated with a HRP conjugated goat-anti-rabbit antibody for 5hmC, 5fC and 5caC (1:3000, Bio-Rad #1706515) or goat-anti-mouse for 5mC (1:3000, Bio-Rad #1706516). Following treatment with Clarity Western ECL substrate (Bio-rad, CA, USA), membranes were scanned individually on a ChemiDoc MP imaging system (Bio-Rad, CA, USA). Raw images were minimally processed using Photoshop: each blot was individually contrast-corrected using 'Auto contrast' and exposure was decreased evenly across all blots according to image standards.

ELISA. 426 bp dsDNA oligos containing 5mC, 5hmC, 5fC, 5cacC or C (GeneTex, CA, USA) was diluted to a concentration of 50ng/mL in coating buffer (1M NaCI, 50 Mm Na_2PO_4, 0.02% (w/v) NaN_3, pH 7.0) then 50µl were placed into each well of black 96-well plates (4titude, UK) and incubated overnight at 37°C. Plates were blocked for 1h at room temperature in Blocker Casein in PBS (Thermofischer Scientific, MA, US) followed by washing with 100 µl PBS containing 0.1% (v/v) Tween 20. Wells were incubated with 50µl of their respective antibodies (1:1000, see above) for 1h at room temperature, then washed 3 times and incubated with 50µl

of horseradish peroxidase (HRP)-conjugated goat-anti- mouse or goat-anti- rabbit antibody (1:5000, see above) for 30 min. Plates were treated with 70µl of Clarity Western ECL substrate (Bio-rad, CA, USA) for 5 min then scanned in a Spark 10M multimode microplate reader (Tecan Trading AG, Switzerland).

Uniform analysis pipeline for processing of published DIP-Seq data. All datasets used are outlined in **Supplementary Table 1**. Raw 5modC DIP-seq sequencing data was downloaded from GSE42250, GSE24841, GSE31343, ERP000570, GSE28500 and GSE55049 then aligned to the mouse genome (mm9) using Bowtie2[45] (bowtie2 -N 1 -L 30). Genomic coverage was calculated using Bedtools[46] (bedtools genomecov -bg -split) then normalized as reads per million mapped (RPM) for visualization where specified. Identification of enriched regions was performed using MACS2[47] (macs2 --bw=200 -p 1e-5) using IgG or Input controls from the same study where possible otherwise IgG or Input samples from the above studies were pooled and randomly subsampled to 20 million reads as controls. Unless otherwise stated, 5modC enriched regions were identified using IgG controls and IgG enriched regions using Input.

6mA DIP-seq data was downloaded from GSE71866, GSE74184, GSE76740 and GSE79543 and processed as 5modC DIP-seq data (see above) except for *X.laevis* data which was aligned to the Refseq Xenopus_laevis_v2 genome (GCF_001663975.1).

Bisulfite sequencing data was obtained from GSE41923 and aligned to a bisulfite converted mm9 index using Bismark[48] (bismark –N 1). Methylation levels of Cytosines in both CpG and non-CpG contexts were extracted for bases with at least 5X coverage (bismark_methylation_extractor –p –comprehensive –bedgraph –buffer_size 75% --cutoff 5).

Raw 5hmC-Seal data was downloaded from GSE41545 and processed as DIP-seq data (see above) and anti-CMS was downloaded from GSE28682 and aligned using Bismark[48] with the same settings as for DIP-seq (bismark -N 1 -L 30).

TET1 ChIP-seq data was downloaded from GSE24843 and histone ChIP-seq data for mESCs was obtained from the ENCODE project[49] and processed as DIP-seq data (see above).

See **Supplementary Table 1** for specification of files used for each analysis/figure.

Analysis of PCR bias. Mapped reads from DIP and Seal techniques were extended to 200 bp to represent sequenced fragments and GC content was counted per "fragment". Theoretical distribution was modelled as a normal distribution after observed data. Molecular complexity in the form of non-redundant read fraction was calculated using Pre-seq[50] (preseq c_curve) at a depth of 10 million reads.

Estimation of number of DIP-Seq studies that include an IgG-Seq control. The Gene Expression Omnibus was searched with the query string, "(meDIP-Seq OR hmeDIP-Seq OR DIP-Seq)", in January 2018. This search returned 153 unique studies, of which 8 were found (by manual curating) to use an IgG-Seq control; 95% of studies did not include an IgG control.

Estimation of falsely enriched regions. Enriched regions were obtained from MACS2 using either pooled IgG or Input from mESCs as control (see above). True positive regions were defined as enriched regions identified for both IgG and Input controls (overlapping regions) and false positive regions were calculated as the inverse fraction of non-overlapping regions for either control. This is visualized in **Supplementary Fig. 3a**.

Motif enrichment of FASTQ files. FASTQ files were trimmed of adapters using ea-utils[51] (fastq-mcf -x 0 -q 0 -k 0 -s 4.6) then randomly subsampled to 1 million reads and subjected to *de novo* motif enrichment analysis using Homer2[52] (homer2 denovo -len 12). Input samples from the same study was used as background when available, otherwise a pooled input from

multiple studies was used (see above). Correlation between motif PWMs was performed using Pearson correlation as implemented in TFBStools[53] (PWMsimilarity), subject motifs were repeated once to account for base shifts. To identify if motifs belong to a certain repeat class, motif PWMs were mapped to repeats in mouse (RepBase v22.01[54]) using Homer2[52] (scanMotifGenomeWide.pl). SRX1141880 was excluded from motif analysis since it contained less than 2 million mapped reads.

Taxonomic annotation of sequence reads. Species classification was performed using Centrifuge[55] (1.0.3-Beta) which is specifically designed for metagenomics classification. Although Centrifuge utilizes similar indexing algorithms as Bowtie2, it far outperforms it for microbial classification[55]. A custom Centrifuge index was built from available complete RefSeq genomes of common cell culture contaminants[56-58], including bacteria, virus and fungi, together with the mouse genome (mm9). The 324 different assemblies included are available in **Supplementary Table 4.**

For determination of short tandem repeat (STR) fraction of species genomes, Tandem Repeat Finder[59] (TRF) results for genomes (ce10, danRer10, dm6, hg38, mm10) was obtained from UCSC. For *X.laevis* and *E.coli* (K-12) the genomic sequence was obtained from Refseq accessions Xenopus_laevis_v2 (GCF_001663975.1) and ASM584v2, respectively, and STRs was identified using TRF 4.09 with recommended settings and a maximum period size of 12 (trf 2 7 7 80 10 50 12).

GO term enrichment analysis. Top 500 enriched regions were mapped to the nearest gene within 10kb and enrichment of GO terms biological processes was performed using PANTHER[60] with default settings.

Statistics and Reproducibility. All statistical analysis was performed using the statistical programming language R[61] unless otherwise stated. P-values <0.05 were considered

significant. All statistical tests were performed as two-tailed unless otherwise stated. Kolmogorov–Smirnov test was used to non-parametrically compare the mean of distributions in **Supplementary Fig. 2h.**

Representative genome browser figures **Fig 1a-b** and **Fig 2a** were reproducible in over 30 biologically independent samples across at least 7 published articles from different groups (see **Supplementary Fig. 1**). The controls experiment in **Supplementary Fig. 2a** was performed once and reproducible in 3 independent experiments using a different method (see **Supplementary Fig. 2b**). Results in **Supplementary Fig. 4h** was reproducible in 26 biologically independent samples from ENCODE (see **Fig. 3e**).

Code availability. Scripts for specific analyses have been deposited to GitHub (https://github.com/ALentini/DIPseqPaper).

Data availability. The sequencing data that supports the findings of this study are publicly available through GEO or ENA under accessions GSE42250[62], GSE24843[63], GSE31343[64], ERP000570[65], GSE28500[66], GSE71866[67], GSE74184[68], GSE76740[69], GSE79543[70], GSE66504[71], GSE55049[72], GSE41923[73], GSE41545[74], GSE28682[75] and mouse ENCODE[49] data is available from https://www.encodeproject.org/.

See **Supplementary Table 1** for specification of files used for each analysis/figure.

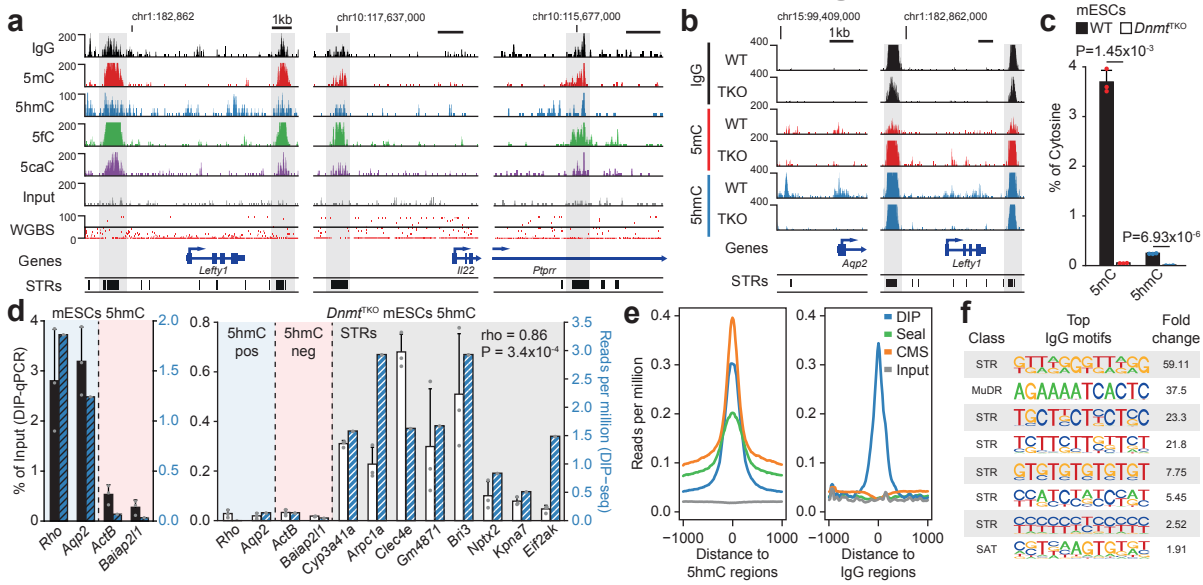
METHODS-ONLY REFERENCES

44. Tsumura, A. et al. Maintenance of self-renewal ability of mouse embryonic stem cells in the absence of DNA methyltransferases Dnmt1, Dnmt3a and Dnmt3b. *Genes to cells : devoted to molecular & cellular mechanisms* **11**, 805-814 (2006).
45. Langmead, B. & Salzberg, S.L. Fast gapped-read alignment with Bowtie 2. *Nature methods* **9**, 357-359 (2012).
46. Quinlan, A.R. & Hall, I.M. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics* **26**, 841-842 (2010).

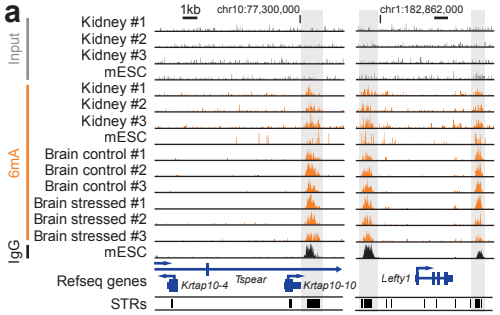
47. Zhang, Y. et al. Model-based analysis of ChIP-Seq (MACS). *Genome biology* **9**, R137 (2008).
48. Krueger, F. & Andrews, S.R. Bismark: a flexible aligner and methylation caller for Bisulfite-Seq applications. *Bioinformatics* **27**, 1571-1572 (2011).
49. Yue, F. et al. A comparative encyclopedia of DNA elements in the mouse genome. *Nature* **515**, 355-364 (2014).
50. Daley, T. & Smith, A.D. Predicting the molecular complexity of sequencing libraries. *Nature methods* **10**, 325-327 (2013).
51. Aronesty, E. Comparison of Sequencing Utility Programs. *The Open Bioinformatics Journal* **7**, 1-8 (2013).
52. Heinz, S. et al. Simple combinations of lineage-determining transcription factors prime cis-regulatory elements required for macrophage and B cell identities. *Molecular cell* **38**, 576-589 (2010).
53. Tan, G. & Lenhard, B. TFBSTools: an R/bioconductor package for transcription factor binding site analysis. *Bioinformatics* **32**, 1555-1556 (2016).
54. Bao, W., Kojima, K.K. & Kohany, O. Repbase Update, a database of repetitive elements in eukaryotic genomes. *Mobile DNA* **6**, 11 (2015).
55. Kim, D., Song, L., Breitwieser, F.P. & Salzberg, S.L. Centrifuge: rapid and sensitive classification of metagenomic sequences. *Genome research* **26**, 1721-1729 (2016).
56. Merten, O.W. Virus contaminations of cell cultures - A biotechnological view. *Cytotechnology* **39**, 91-116 (2002).
57. Drexler, H.G. & Uphoff, C.C. Mycoplasma contamination of cell cultures: Incidence, sources, effects, detection, elimination, prevention. *Cytotechnology* **39**, 75-90 (2002).
58. Ali, S. Microbial and Viral Contamination of Animal and Stem Cell Cultures: Common Contaminants, Detection and Elimination. *Journal of Stem Cell Research & Therapeutics* **2** (2017).
59. Benson, G. Tandem repeats finder: a program to analyze DNA sequences. *Nucleic acids research* **27**, 573-580 (1999).
60. Mi, H., Muruganujan, A. & Thomas, P.D. PANTHER in 2013: modeling the evolution of gene function, and other gene attributes, in the context of phylogenetic trees. *Nucleic acids research* **41**, D377-386 (2013).
61. R Development Core Team R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0. *http://www.R-project.org* (2008).
62. Shen, L. et al. Genome-wide analysis reveals TET- and TDG-dependent 5-methylcytosine oxidation dynamics. *Cell* **153**, 692-706 (2013).
63. Williams, K. et al. TET1 and hydroxymethylcytosine in transcription and DNA methylation fidelity. *Nature* **473**, 343-348 (2011).
64. Matarese, F., Carrillo-de Santa Pau, E. & Stunnenberg, H.G. 5-Hydroxymethylcytosine: a new kid on the epigenetic block? *Molecular systems biology* **7**, 562 (2011).
65. Ficz, G. et al. Dynamic regulation of 5-hydroxymethylcytosine in mouse ES cells and during differentiation. *Nature* **473**, 398-402 (2011).
66. Xu, Y. et al. Genome-wide regulation of 5hmC, 5mC, and gene expression by Tet1 hydroxylase in mouse embryonic stem cells. *Molecular cell* **42**, 451-464 (2011).
67. Wu, T.P. et al. DNA methylation on N(6)-adenine in mammalian embryonic stem cells. *Nature* **532**, 329-333 (2016).
68. Koziol, M.J. et al. Identification of methylated deoxyadenosines in vertebrates reveals diversity in DNA modifications. *Nature structural & molecular biology* **23**, 24-30 (2016).

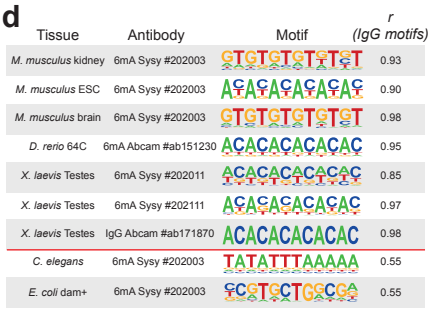
69. Liu, J. et al. Abundant DNA 6mA methylation during early embryogenesis of zebrafish and pig. *Nature communications* **7**, 13052 (2016).
70. Yao, B. et al. DNA N6-methyladenine is dynamically regulated in the mouse brain following environmental stress. *Nature communications* **8**, 1122 (2017).
71. Greer, E.L. et al. DNA Methylation on N6-Adenine in C. elegans. *Cell* **161**, 868-878 (2015).
72. Dawlaty, M.M. et al. Loss of Tet enzymes compromises proper differentiation of embryonic stem cells. *Developmental cell* **29**, 102-111 (2014).
73. Habibi, E. et al. Whole-genome bisulfite sequencing of two distinct interconvertible DNA methylomes of mouse embryonic stem cells. *Cell stem cell* **13**, 360-369 (2013).
74. Song, C.X. et al. Genome-wide profiling of 5-formylcytosine reveals its roles in epigenetic priming. *Cell* **153**, 678-691 (2013).
75. Pastor, W.A. et al. Genome-wide mapping of 5-hydroxymethylcytosine in embryonic stem cells. *Nature* **473**, 394-397 (2011).

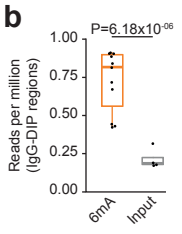
Lentini et al. Figure 1

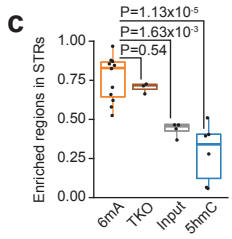


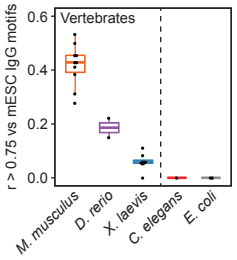
Lentini et al. Figure 2

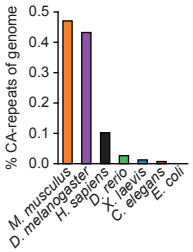




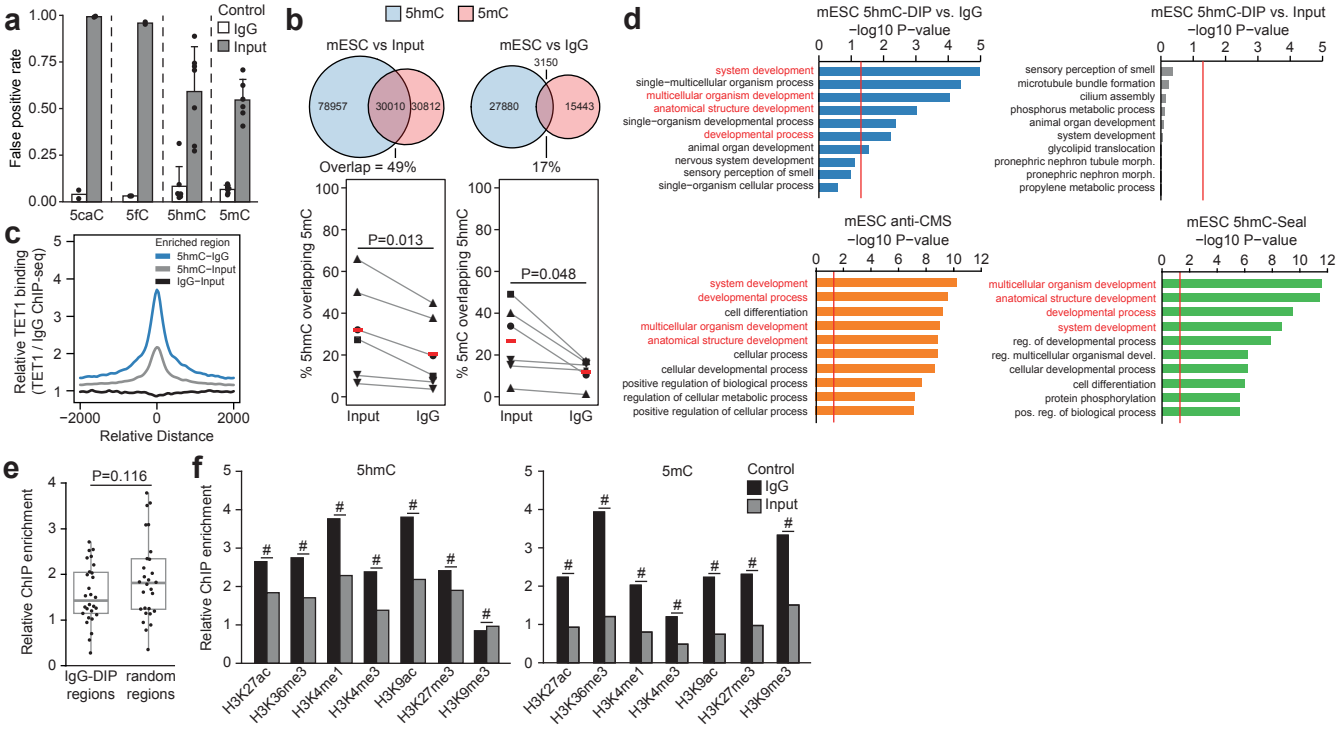




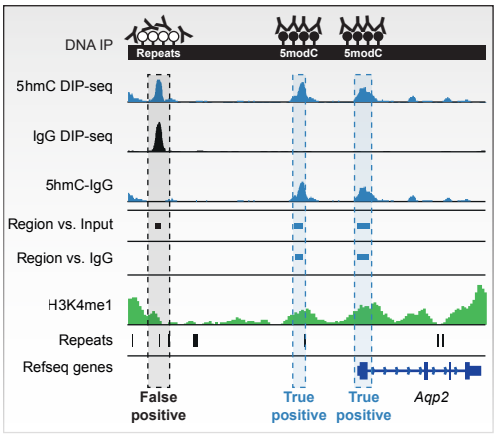




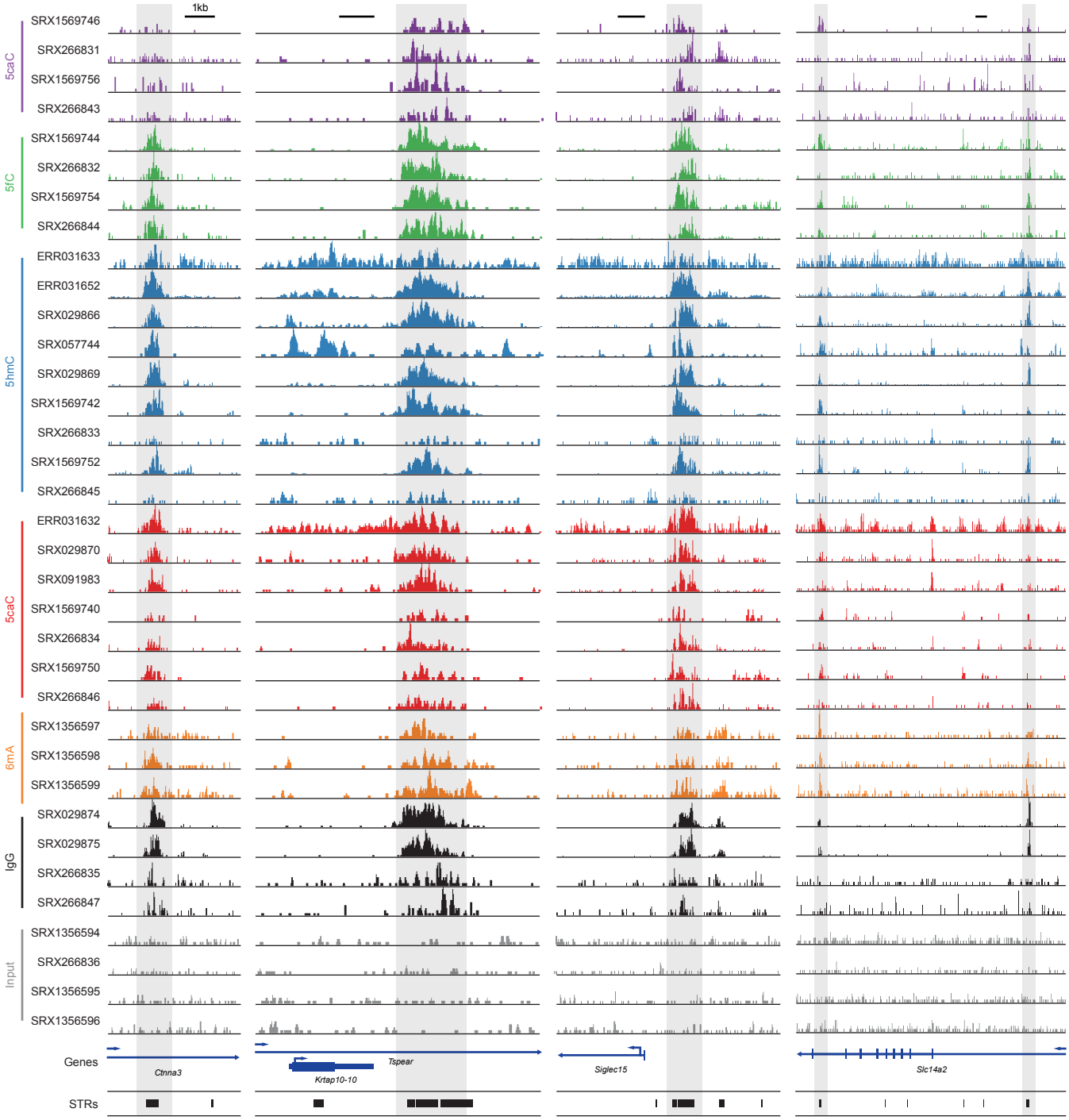
Lentini et al. Figure 3



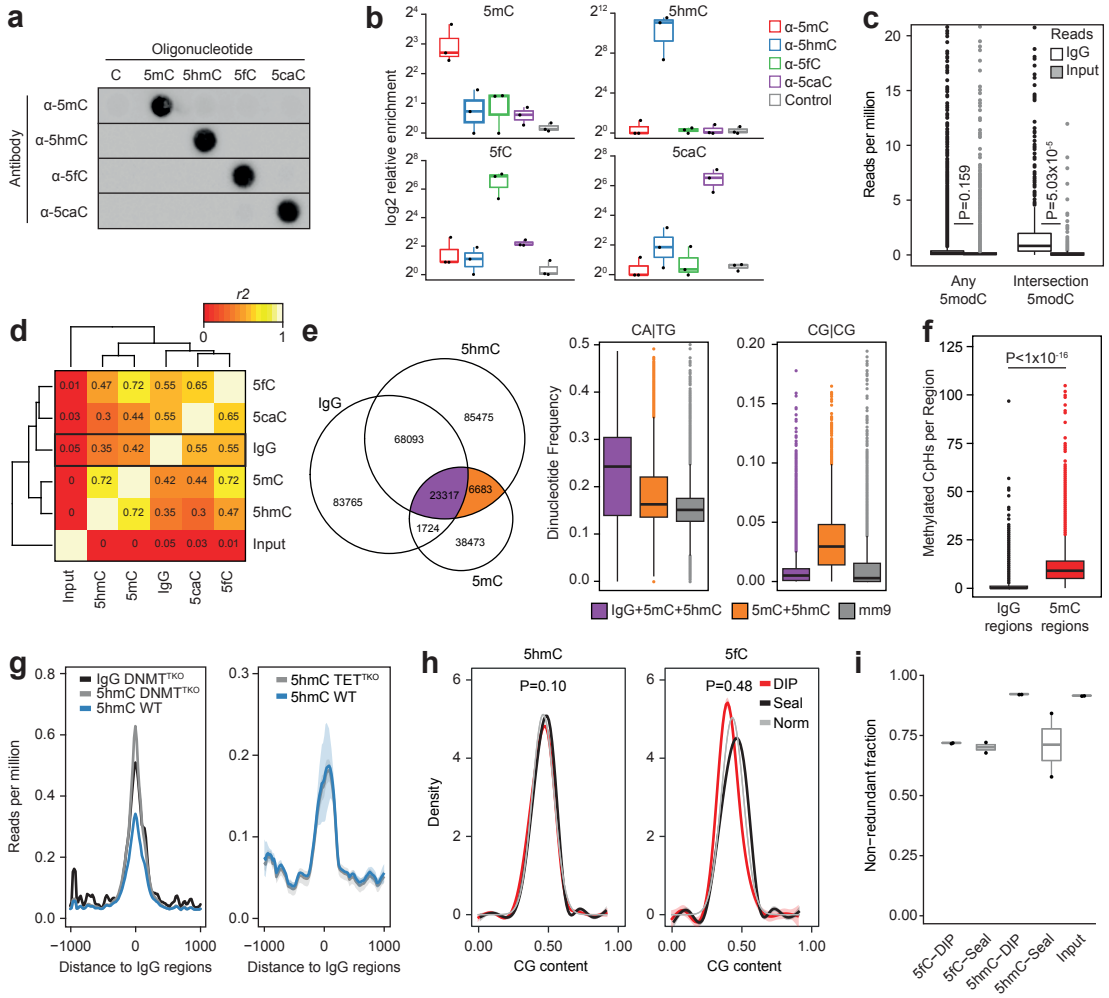
Lentini et al. Figure 4



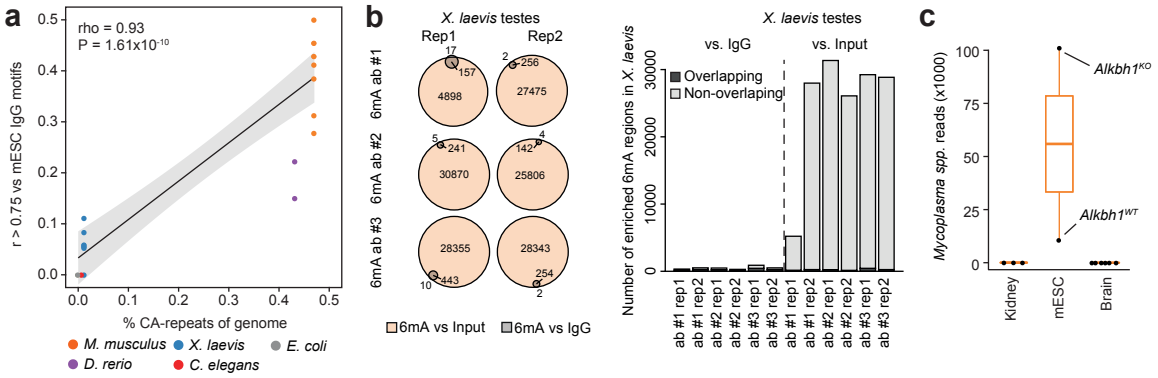
Lentini et al. Supplementary Figure 1



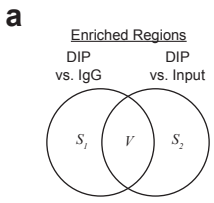
Lentini et al. Supplementary Figure 2



Lentini et al. Supplementary Figure 3



Lentini et al. Supplementary Figure 4



False Positive Rate:

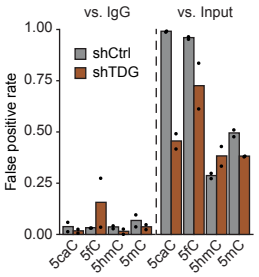
$$FPR_{IgG} = 1 - \frac{V}{(S_1 + V)}$$

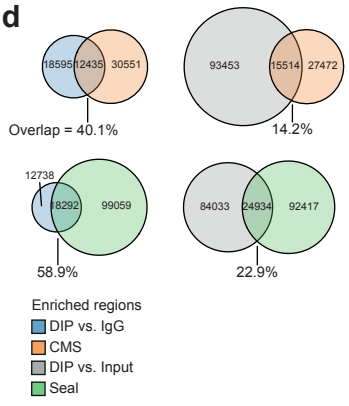
$$FPR_{Input} = 1 - \frac{V}{(S_2 + V)}$$

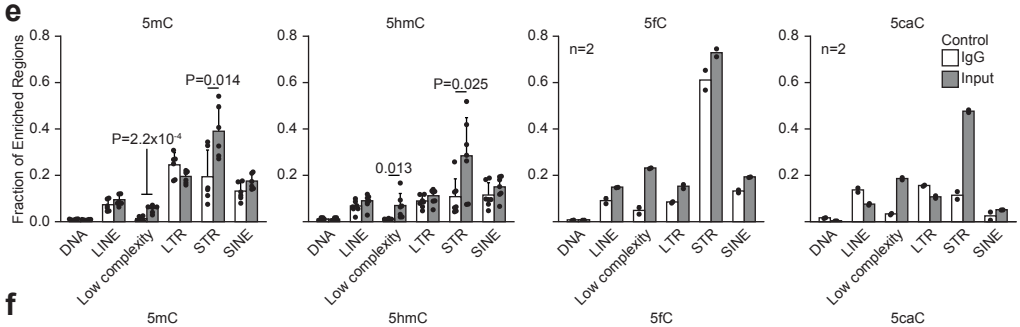
Where V = True Positives

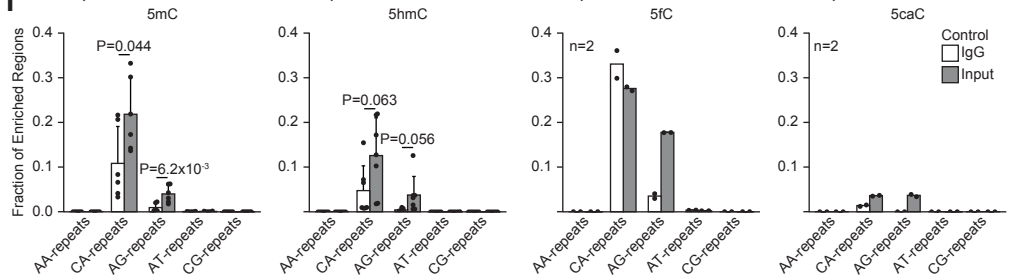
	Ref	Mark	vs. IgG	vs. Input
mESC	ERP000570	5hmC rep1	0.315	0.890
		5hmC rep2	0.025	0.745
		5mC rep1	0.073	0.627
		5mC rep2	0.068	0.552
	GSE31343	5hmC	0.094	0.708
		5mC	0.081	0.408
	GSE42250	5caC rep1	0.061	0.989
		5caC rep2	0.015	0.994
		5fC rep1	0.033	0.954
		5fC rep2	0.033	0.966
		5hmC rep1	0.030	0.300
		5hmC rep2	0.044	0.275
		5mC rep1	0.098	0.479
		5mC rep2	0.039	0.512
	GSE24841	5hmC	0.045	0.728
		5mC	0.048	0.709
	GSE28500	5hmC	0.041	0.506
MEF	GSE66438*	5caC rep1	0.090	0.314
		5caC rep2	0.081	0.258
		5fC rep1	0.027	0.668
		5fC rep2	0.014	0.376
		5hmC rep1	0.037	0.307
		5hmC rep2	0.025	0.441
		5mC rep1	0.045	0.141
		5mC rep2	0.052	0.173
Kidney	GSE74184*	6mA rep1	0.243	0.669
		6mA rep2	0.280	0.712
		6mA rep3	0.241	0.725
mESC	GSE71866	6mA KO^{ALKBH1}	0.054	0.372
		6mA WT^{ALKBH1}	0.009	0.172

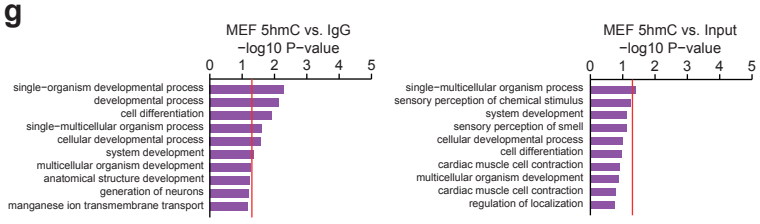
*Approximated using mESC control data

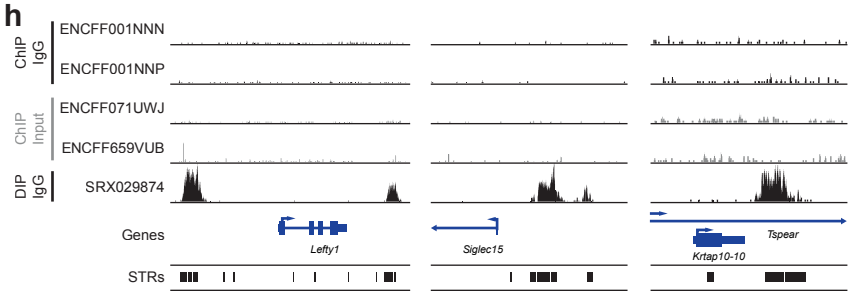












Supplementary Discussion

Normalizing for off-target binding in DIP-Seq

The prevalence of non-enriched Input DNA as a control in DIP-Seq studies stems from its use in ChIP-seq; Input chromatin helps to control for the different shearing dynamics of closed and open chromatin and for differences in the amplification efficiency of DNA fragments with different base compositions[1]. The preference for Input controls was also fueled by the requirement of a uniform background signal in early peak-calling algorithms[2]. Furthermore, comparison between Input and a control antibody in ChIP-seq has shown negligible differences[3] but such a comparison has, to our knowledge, never been performed for DIP-seq until now. While Input controls for sequencing bias and genome mappability, it does not correct for antibody cross-reactivity and subsequently introduces genome-wide biases in the data. We show that error-rate of off-target binding is highly dependent on the mark of interest and may account for 50-99% of the observed enrichment whereas error-rates related to mappability is consistent across targets at around 5-6%. Due to the large disparity between controls, IgG should be used as a control as it allows consistent background removal and minimizes errors. It is noteworthy that comparative studies utilizing biological controls (such as knockouts) have been less affected by these errors[4-6] but this is not possible for novel modifications without known enzymatic pathways. It is also important to appreciate that nearly 80% of genes in mice (mm9) contain STRs that may act as functional regulators[4] making masking procedures such as blacklisting ill-advised. Thus, we strongly suggest that all future DIP-seq studies perform both Input and IgG controls. This also stresses the importance of independent validation of findings. Currently DIP-qPCR is commonly used for experimental validation but still suffers by antibody cross-reactivity (**Fig. 1d**). Other techniques such as bisulfite sequencing (BS), methyl-sensitive restriction enzyme digestion and non-antibody based enrichment techniques represents complementary methodology that should be considered[7]. Indeed, future profiling studies of DNA modifications may be advised to use non-antibody based mapping techniques where possible[7]. Bisulfite sequencing of 5mC and oxidative BS or TAB-seq of 5hmC offer quantitative, base-resolution alternatives to DIP-seq, but remain prohibitively expensive[8, 9]. The click chemistry based assays 5hmC-Seal and 5fC-Seal are low-cost enrichment based techniques that do not exhibit STR enrichment bias but may be less sensitive than their antibody-based counterparts[10-12]

Whereas normalization of DIP-seq data to an IgG-seq control represents the optimal approach to generating accurate DIP-seq profiles, IgG controls are lacking for the majority of published studies. Computational correction of published DIP-seq data by filtering out sequencing reads containing IgG associated STR motifs is relatively straightforward, but is not advised. First, as DNA modifications (5mC, 5hmC, 5fC, 5caC) do occur at non-CpG dinucleotides in some cell types, complete removal of IgG-STR sequences may result in a loss of biologically significant information[4, 13] (**Supplementary Fig. 4e,f**). Second, as genomic STR composition differs markedly between species, the set of STRs bound by IgG and the extent of their enrichment is likely to vary in DIP-seq of DNA from different organisms. Third, as the effect of off-target STR binding increases with decreasing abundance of the target epitope (**Fig. S4b**), *a priori* knowledge of global modification levels in each genome and cell type would be required to prevent over-correction of the data. Finally, other experimental variables such as antibody source and sensitivity, DNA denaturation conditions and stringency of washing may also effect the degree of STR-binding observed. Consequently, optimal reanalysis of published DIP-seq data requires the generation of additional IgG-seq data for each cell type under investigation.

References

1. Kidder, B.L., Hu, G. & Zhao, K. ChIP-Seq: technical considerations for obtaining high-quality data. *Nature immunology* **12**, 918-922 (2011).
2. Yao, B. et al. DNA N6-methyladenine is dynamically regulated in the mouse brain following environmental stress. *Nature communications* **8**, 1122 (2017).
3. Flensburg, C., Kinkel, S.A., Keniry, A., Blewitt, M.E. & Oshlack, A. A comparison of control samples for ChIP-seq of histone modifications. *Frontiers in genetics* **5**, 329 (2014).
4. Papin, C. et al. Combinatorial DNA methylation codes at repetitive elements. *Genome research* **27**, 934-946 (2017).
5. Shen, L. et al. Genome-wide analysis reveals TET- and TDG-dependent 5-methylcytosine oxidation dynamics. *Cell* **153**, 692-706 (2013).
6. Williams, K. et al. TET1 and hydroxymethylcytosine in transcription and DNA methylation fidelity. *Nature* **473**, 343-348 (2011).
7. Nestor, C.E., Reddington, J.P., Benson, M. & Meehan, R.R. Investigating 5-hydroxymethylcytosine (5hmC): the state of the art. *Methods in molecular biology* **1094**, 243-258 (2014).
8. Booth, M.J. et al. Quantitative sequencing of 5-methylcytosine and 5-hydroxymethylcytosine at single-base resolution. *Science* **336**, 934-937 (2012).
9. Yu, M. et al. Base-resolution analysis of 5-hydroxymethylcytosine in the mammalian genome. *Cell* **149**, 1368-1380 (2012).
10. Thomson, J.P. et al. Comparative analysis of affinity-based 5-hydroxymethylation enrichment techniques. *Nucleic acids research* **41**, e206 (2013).
11. Song, C.X. et al. Genome-wide profiling of 5-formylcytosine reveals its roles in epigenetic priming. *Cell* **153**, 678-691 (2013).
12. Song, C.X. et al. Selective chemical labeling reveals the genome-wide distribution of 5-hydroxymethylcytosine. *Nat Biotechnol* **29**, 68-72 (2011).
13. Habibi, E. et al. Whole-genome bisulfite sequencing of two distinct interconvertible DNA methylomes of mouse embryonic stem cells. *Cell Stem Cell* **13**, 360-369 (2013).

SUPPLEMENTARY METHODS

Cell culture. J1 mouse embryonic stem cells (mESCs; WT, male) were originally derived from the 129S4/SvJae strain. TKO (Dnmt1-/-, Dnmt3a-/-, Dnmt3b-/-) mESCs were derived from J1 mESCs[44]. Both cell lines were cultured in a humidified incubator at 5% CO_2, 37°C on 0.2% gelatin coated tissue culture plastic in DMEM (Dulbecco's modified eagle medium) supplemented with 15 % fetal calf serum, 0.1 mM non-essential amino acids (Sigma-Aldrich, MI, USA), 1 mM sodium Pyruvate (Sigma-Aldrich, MI, USA), 1 % Penicillin/Streptomycin, 2 mM L-glutamine, 0.1 mM beta-mercaptoethanol (Thermo Fisher, CA, USA), and ESGRO LIF (Millipore, MA, USA) at 500U/mL. mESCs were passaged every 2-3 days using trypsin/EDTA.

DNA extraction. Snap frozen cell pellets were treated with RNAse cocktail (Ambion, CA, USA) for 1 hour at 37°C followed by proteinase K treatment overnight at 55°C. DNA was extracted by standard phenol chloroform/ethanol precipitation and eluted in TE.

DIP-qPCR. 1.5 µg genomic DNA was sonicated to fragments ranging between 100-1000 bp with a peak at 400 bp using a BioRuptor (Diagenode, Belgium), denatured at 95°C for 10 min then cooled on wet ice for 10 min. 10% of samples were saved as Input and the remaining DNA was resuspended in 10x IP buffer (10 mM Na-Phosphate (mono-dibasic), 1% NaCl, 0.05% Triton X-100, pH 7.0). Immunoprecipitations were performed using 1µg anti-5hmC antibody (Active Motif, #39769) for 12h at 4°C using constant rotation. Protein G dynabeads (Invitrogen, CA, USA, #100-03D) were washed twice in 0.1% PBS-BSA then added to the IP mixture for 1h at 4° using constant rotation. Beads were washed three times for 10 min using cold 1x IP buffer then resuspended in digestion buffer and incubated with 8 U Proteinase K (New England Biolabs, MA, USA) for 1.5h at 50°C, 800rpm in 50 mM Tris, 10 mM EDTA 0.5% SDS, pH 8.0 and purified using DNA Clean & Concentrator kit (Zymo Research, USA).

Quantitative PCR was performed on a 7900HT real-time cycler (Applied Biosystems, CA, USA) using SYBR green master mix (Applied Biosystems, CA, USA). qPCR primers use are listed in **Supplementary Table 4**, below.

Supplementary Table 4. hMeDIP qPCR primer sequences

name	forward primer (5' - 3')	reverse primer (5' – 3')	designation
Rho	ACCGTACAGCACAAGAAGCTGC	GAAGACCATGAAGAGGTCAGCC	True Positive
Aqp2	ATGTGGGAACTCCGGTCCATAG	GCCAAAGAAGACGAAAAGGAGC	True Positive
ActB	ATGAAGAGTTTTGGCGATGG	GATGCTGACCCTCATCCACT	True Negative
Baiap2l1	ATCTGCACTTGATGACAACTGG	CTTGTGAGACCAAGCTCTTAGC	True Negative
Cyp3a41a	TTCACCTTTATGACTTGGTAGGC	GCTTCTCTTGTGAGGACTGTGG	False Positive
Arpc1a	TGGGGCTCATTTCTGTAATACC	TTCCATCTTCTCAAATCATTGC	False Positive
Nptx2	TCTCAAGTGCTGGGATTAAAGG	TCTGGGAAGCAAATCTAAGTCC	False Positive
Gm4871	CTGGTGTGTGTTTATCCTCAGC	AACTGTGGAGTGAGGTATGAAGG	False Positive
Bri3	TGGAGAGTGTGTATGTGTGAGC	AGGAGGCAGAAGGAGAAAGG	False Positive
Clec4e	CACATACTGCCTTCTGCTATGC	TGTGTGAGTGAAAGGAGAGAGC	False Positive
Kpna7	CAACCAGGACTACACAGTGACG	GACACAGAAGCACAGAGAGAGG	False Positive
Eif2ak	AGAGGCCAGAAGGTGTTGG	TTTCAGAGGACCTGAGTTTGG	False Positive

Quantification of cytosine modifications using mass spectrometry. 1 µg of DNA was heat denatured at 100 °C for 5 min in 20µL H_2O then immediately cooled on ice. 10 µl P1 Nuclease (0.02 U/µl in 90 mM AmAc, 0.3 mM $ZnSO_4$, pH 5.3) was added followed by incubation at 50 °C for 2 h. 10 µl Alkaline phosphatase (0.08 U/µl in 200 mM TRIS-HCl, 0.40 mM EDTA, pH 8) was added followed by incubation at 37 °C for 30 min. Proteins were precipitated by the addition of 160 µl cold acetonitrile. Following centrifugation at 17000 x g for 5 min, 180 µl of the supernatant was evaporated under nitrogen and reconstituted in 40 µl 0.1% formic acid. The chromatographic system consisted of an Acquity UPLC (Waters, MA, USA) and a Xevo triple quadrupole mass spectrometer (Waters, MA, USA). The extracts were separated on an HSS T3 column (150x2.1 mm, 1.7 µm, Waters, MA, USA) at 45°C and a flow rate of 450 µl/min using a gradient elution with 0.05% acetic acid and methanol, 0-1.3 min 2% B; 1.3-5.5

min 2-9% B; 5.5-7.5 min re-equilibration at 2% B. For dC a 1 µl injection was made and for mC, hmC, fC and caC a 15 µl injection was made. Analytes were detected in the multi reaction monitoring (MRM) mode using three time windows with the following transistions 0-2.3 min – C (228->95 & 228->112) and hmC (258->124 & 258->142); 2.3-4 min – mC (242->109, 242->126) and caC(272->138, 272->156); 4-7.5 min – fC (256->97, 256->140).

Immuno dot-blot. 10 ng 426 bp oligos containing 5mC, 5hmC, 5fC, 5caC or C (GeneTex, CA, USA) was denatured at 95°C for 15 min in 0.4M NaOH and 10mM EDTA then immediately cooled on ice. Samples were applied to a positively charged nylon membrane under vacuum using a Dot Blot Hybridisation Manifold (Harvard Apparatus, MA, USA). The membranes were briefly washed in 2X SSC buffer (0.3M NaCl, 30mM NaCitrate) then cross-linked using a UV Stratalinker 1800 (Stratagene, CA, USA) and baked at 80°C for 2 h. Membranes were blocked in casein blocking buffer (Li-Cor) for 15 min at 4°C then incubated with an antibody against 5mC (1:3000, Zymo #A3001), 5hmC (1:3000, ActiveMotif #39791), 5fC (1:3000, ActiveMotif #61227) or 5caC (1:3000, ActiveMotif #61229) for 1h at 4°C. Membranes were washed 3 times for 5 min in TBS-Tween (0.05%) then incubated with a HRP conjugated goat-anti-rabbit antibody for 5hmC, 5fC and 5caC (1:3000, Bio-Rad #1706515) or goat-anti-mouse for 5mC (1:3000, Bio-Rad #1706516). Following treatment with Clarity Western ECL substrate (Bio-rad, CA, USA), membranes were scanned individually on a ChemiDoc MP imaging system (Bio-Rad, CA, USA). Raw images were minimally processed using Photoshop: each blot was individually contrast-corrected using 'Auto contrast' and exposure was decreased evenly across all blots according to image standards.

ELISA. 426 bp dsDNA oligos containing 5mC, 5hmC, 5fC, 5cacC or C (GeneTex, CA, USA) was diluted to a concentration of 50ng/mL in coating buffer (1M NaCI, 50 Mm Na_2PO_4, 0.02% (w/v) NaN_3, pH 7.0) then 50µl were placed into each well of black 96-well plates (4titude, UK) and incubated overnight at 37°C. Plates were blocked for 1h at room temperature in Blocker

Casein in PBS (Thermofischer Scientific, MA, US) followed by washing with 100 µl PBS containing 0.1% (v/v) Tween 20. Wells were incubated with 50µl of their respective antibodies (1:1000, see above) for 1h at room temperature, then washed 3 times and incubated with 50µl of horseradish peroxidase (HRP)-conjugated goat-anti- mouse or goat-anti- rabbit antibody (1:5000, see above) for 30 min. Plates were treated with 70µl of Clarity Western ECL substrate (Bio-rad, CA, USA) for 5 min then scanned in a Spark 10M multimode microplate reader (Tecan Trading AG, Switzerland).

Uniform analysis pipeline for processing of published DIP-Seq data. All datasets used are outlined in **Supplementary Table 1**. Raw 5modC DIP-seq sequencing data was downloaded from GSE42250, GSE24841, GSE31343, ERP000570, GSE28500 and GSE55049 then aligned to the mouse genome (mm9) using Bowtie2[45] (bowtie2 -N 1 -L 30). Genomic coverage was calculated using Bedtools[46] (bedtools genomecov -bg -split) then normalized as reads per million mapped (RPM) for visualization where specified. Identification of enriched regions was performed using MACS2[47] (macs2 --bw=200 -p 1e-5) using IgG or Input controls from the same study where possible otherwise IgG or Input samples from the above studies were pooled and randomly subsampled to 20 million reads as controls. Unless otherwise stated, 5modC enriched regions were identified using IgG controls and IgG enriched regions using Input.

6mA DIP-seq data was downloaded from GSE71866, GSE74184, GSE76740 and GSE79543 and processed as 5modC DIP-seq data (see above) except for *X.laevis* data which was aligned to the Refseq Xenopus_laevis_v2 genome (GCF_001663975.1).

Bisulfite sequencing data was obtained from GSE41923 and aligned to a bisulfite converted mm9 index using Bismark[48] (bismark –N 1). Methylation levels of Cytosines in both CpG and

non-CpG contexts were extracted for bases with at least 5X coverage (bismark_methylation_extractor –p –comprehensive –bedgraph –buffer_size 75% --cutoff 5).

Raw 5hmC-Seal data was downloaded from GSE41545 and processed as DIP-seq data (see above) and anti-CMS was downloaded from GSE28682 and aligned using Bismark[48] with the same settings as for DIP-seq (bismark -N 1 -L 30).

TET1 ChIP-seq data was downloaded from GSE24843 and histone ChIP-seq data for mESCs was obtained from the ENCODE project[49] and processed as DIP-seq data (see above).

See **Supplementary Table 1** for specification of files used for each analysis/figure.

Analysis of PCR bias. Mapped reads from DIP and Seal techniques were extended to 200 bp to represent sequenced fragments and GC content was counted per "fragment". Theoretical distribution was modelled as a normal distribution after observed data. Molecular complexity in the form of non-redundant read fraction was calculated using Pre-seq[50] (preseq c_curve) at a depth of 10 million reads.

Estimation of number of DIP-Seq studies that include an IgG-Seq control. The Gene Expression Omnibus was searched with the query string, "(meDIP-Seq OR hmeDIP-Seq OR DIP-Seq)", in January 2018. This search returned 153 unique studies, of which 8 were found (by manual curating) to use an IgG-Seq control; 95% of studies did not include an IgG control.

Estimation of falsely enriched regions. Enriched regions were obtained from MACS2 using either pooled IgG or Input from mESCs as control (see above). True positive regions were defined as enriched regions identified for both IgG and Input controls (overlapping regions) and false positive regions were calculated as the inverse fraction of non-overlapping regions for either control. This is visualized in **Supplementary Fig. 3a**.

Motif enrichment of FASTQ files. FASTQ files were trimmed of adapters using ea-utils[51] (fastq-mcf -x 0 -q 0 -k 0 -s 4.6) then randomly subsampled to 1 million reads and subjected to *de novo* motif enrichment analysis using Homer2[52] (homer2 denovo -len 12). Input samples from the same study was used as background when available, otherwise a pooled input from multiple studies was used (see above). Correlation between motif PWMs was performed using Pearson correlation as implemented in TFBStools[53] (PWMsimilarity), subject motifs were repeated once to account for base shifts. To identify if motifs belong to a certain repeat class, motif PWMs were mapped to repeats in mouse (RepBase v22.01[54]) using Homer2[52] (scanMotifGenomeWide.pl). SRX1141880 was excluded from motif analysis since it contained less than 2 million mapped reads.

Taxonomic annotation of sequence reads. Species classification was performed using Centrifuge[55] (1.0.3-Beta) which is specifically designed for metagenomics classification. Although Centrifuge utilizes similar indexing algorithms as Bowtie2, it far outperforms it for microbial classification[55]. A custom Centrifuge index was built from available complete RefSeq genomes of common cell culture contaminants[56-58], including bacteria, virus and fungi, together with the mouse genome (mm9). The 324 different assemblies included are available in **Supplementary Table 4.**

For determination of short tandem repeat (STR) fraction of species genomes, Tandem Repeat Finder[59] (TRF) results for genomes (ce10, danRer10, dm6, hg38, mm10) was obtained from UCSC. For *X.laevis* and *E.coli* (K-12) the genomic sequence was obtained from Refseq accessions Xenopus_laevis_v2 (GCF_001663975.1) and ASM584v2, respectively, and STRs was identified using TRF 4.09 with recommended settings and a maximum period size of 12 (trf 2 7 7 80 10 50 12).

GO term enrichment analysis. Top 500 enriched regions were mapped to the nearest gene within 10kb and enrichment of GO terms biological processes was performed using PANTHER[60] with default settings.

Statistics and Reproducibility. All statistical analysis was performed using the statistical programming language R[61] unless otherwise stated. P-values <0.05 were considered significant. All statistical tests were performed as two-tailed unless otherwise stated. Kolmogorov–Smirnov test was used to non-parametrically compare the mean of distributions in **Supplementary Fig. 2h.**

Representative genome browser figures **Fig 1a-b** and **Fig 2a** were reproducible in over 30 biologically independent samples across at least 7 published articles from different groups (see **Supplementary Fig. 1**). The controls experiment in **Supplementary Fig. 2a** was performed once and reproducible in 3 independent experiments using a different method (see **Supplementary Fig. 2b**). Results in **Supplementary Fig. 4h** was reproducible in 26 biologically independent samples from ENCODE (see **Fig. 3e**).

Code availability. Scripts for specific analyses have been deposited to GitHub (https://github.com/ALentini/DIPseqPaper).

Data availability. The sequencing data that supports the findings of this study are publicly available through GEO or ENA under accessions GSE42250[62], GSE24843[63], GSE31343[64], ERP000570[65], GSE28500[66], GSE71866[67], GSE74184[68], GSE76740[69], GSE79543[70], GSE66504[71], GSE55049[72], GSE41923[73], GSE41545[74], GSE28682[75] and mouse ENCODE[49] data is available from https://www.encodeproject.org/.

See **Supplementary Table 1** for specification of files used for each analysis/figure.

References

44. Tsumura, A. et al. Maintenance of self-renewal ability of mouse embryonic stem cells in the absence of DNA methyltransferases Dnmt1, Dnmt3a and Dnmt3b. *Genes to cells : devoted to molecular & cellular mechanisms* **11**, 805-814 (2006).
45. Langmead, B. & Salzberg, S.L. Fast gapped-read alignment with Bowtie 2. *Nature methods* **9**, 357-359 (2012).
46. Quinlan, A.R. & Hall, I.M. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics* **26**, 841-842 (2010).
47. Zhang, Y. et al. Model-based analysis of ChIP-Seq (MACS). *Genome biology* **9**, R137 (2008).
48. Krueger, F. & Andrews, S.R. Bismark: a flexible aligner and methylation caller for Bisulfite-Seq applications. *Bioinformatics* **27**, 1571-1572 (2011).
49. Yue, F. et al. A comparative encyclopedia of DNA elements in the mouse genome. *Nature* **515**, 355-364 (2014).
50. Daley, T. & Smith, A.D. Predicting the molecular complexity of sequencing libraries. *Nature methods* **10**, 325-327 (2013).
51. Aronesty, E. Comparison of Sequencing Utility Programs. *The Open Bioinformatics Journal* **7**, 1-8 (2013).
52. Heinz, S. et al. Simple combinations of lineage-determining transcription factors prime cis-regulatory elements required for macrophage and B cell identities. *Molecular cell* **38**, 576-589 (2010).
53. Tan, G. & Lenhard, B. TFBSTools: an R/bioconductor package for transcription factor binding site analysis. *Bioinformatics* **32**, 1555-1556 (2016).
54. Bao, W., Kojima, K.K. & Kohany, O. Repbase Update, a database of repetitive elements in eukaryotic genomes. *Mobile DNA* **6**, 11 (2015).
55. Kim, D., Song, L., Breitwieser, F.P. & Salzberg, S.L. Centrifuge: rapid and sensitive classification of metagenomic sequences. *Genome research* **26**, 1721-1729 (2016).
56. Merten, O.W. Virus contaminations of cell cultures - A biotechnological view. *Cytotechnology* **39**, 91-116 (2002).
57. Drexler, H.G. & Uphoff, C.C. Mycoplasma contamination of cell cultures: Incidence, sources, effects, detection, elimination, prevention. *Cytotechnology* **39**, 75-90 (2002).
58. Ali, S. Microbial and Viral Contamination of Animal and Stem Cell Cultures: Common Contaminants, Detection and Elimination. *Journal of Stem Cell Research & Therapeutics* **2** (2017).
59. Benson, G. Tandem repeats finder: a program to analyze DNA sequences. *Nucleic acids research* **27**, 573-580 (1999).
60. Mi, H., Muruganujan, A. & Thomas, P.D. PANTHER in 2013: modeling the evolution of gene function, and other gene attributes, in the context of phylogenetic trees. *Nucleic acids research* **41**, D377-386 (2013).
61. R Development Core Team R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0. *http://www.R-project.org* (2008).
62. Shen, L. et al. Genome-wide analysis reveals TET- and TDG-dependent 5-methylcytosine oxidation dynamics. *Cell* **153**, 692-706 (2013).
63. Williams, K. et al. TET1 and hydroxymethylcytosine in transcription and DNA methylation fidelity. *Nature* **473**, 343-348 (2011).

64. Matarese, F., Carrillo-de Santa Pau, E. & Stunnenberg, H.G. 5-Hydroxymethylcytosine: a new kid on the epigenetic block? *Molecular systems biology* **7**, 562 (2011).
65. Ficz, G. et al. Dynamic regulation of 5-hydroxymethylcytosine in mouse ES cells and during differentiation. *Nature* **473**, 398-402 (2011).
66. Xu, Y. et al. Genome-wide regulation of 5hmC, 5mC, and gene expression by Tet1 hydroxylase in mouse embryonic stem cells. *Molecular cell* **42**, 451-464 (2011).
67. Wu, T.P. et al. DNA methylation on N(6)-adenine in mammalian embryonic stem cells. *Nature* **532**, 329-333 (2016).
68. Koziol, M.J. et al. Identification of methylated deoxyadenosines in vertebrates reveals diversity in DNA modifications. *Nature structural & molecular biology* **23**, 24-30 (2016).
69. Liu, J. et al. Abundant DNA 6mA methylation during early embryogenesis of zebrafish and pig. *Nature communications* **7**, 13052 (2016).
70. Yao, B. et al. DNA N6-methyladenine is dynamically regulated in the mouse brain following environmental stress. *Nature communications* **8**, 1122 (2017).
71. Greer, E.L. et al. DNA Methylation on N6-Adenine in C. elegans. *Cell* **161**, 868-878 (2015).
72. Dawlaty, M.M. et al. Loss of Tet enzymes compromises proper differentiation of embryonic stem cells. *Developmental cell* **29**, 102-111 (2014).
73. Habibi, E. et al. Whole-genome bisulfite sequencing of two distinct interconvertible DNA methylomes of mouse embryonic stem cells. *Cell stem cell* **13**, 360-369 (2013).
74. Song, C.X. et al. Genome-wide profiling of 5-formylcytosine reveals its roles in epigenetic priming. *Cell* **153**, 678-691 (2013).
75. Pastor, W.A. et al. Genome-wide mapping of 5-hydroxymethylcytosine in embryonic stem cells. *Nature* **473**, 394-397 (2011).