Spring 2021 – Epigenetics and Systems Biology
Discussion Session (Epigenetics)
Michael K. Skinner – Biol 476/576
Week 7 (March 4, 2021)

Epigenetics (History / Molecular Processes / Genomics)

Primary Papers

1. Yao B, et al, (2018) Mol Cell. 2018 Sep 6;71(5):848-857.e6. (PMID: 30078725)
2. Booth, et al. (2012) Science 336:934. (PMID: 22539555)
3. Kelsey, et al. (2017) Science 358:69. (PMID: 28983045)

Discussion

Student 16 – Ref #1 above

- What epigenetic mark was identified?
- What was the technology used?
- What function does the epigenetic mark have?

Student 17 – Ref #2 above

- What is hydroxymethylcytosine and how distinct from 5mC?
- What technology was used?
- What is the function of 5hmC and where expressed?

Student 18 – Ref #3 above

- What epigenetic marks were identified?
- What technology was used?
- How did the genomic profiling correlate with cellular differentiation?

CellPress

Molecular Cell
Short Article

Active N^6-Methyladenine Demethylation by DMAD Regulates Gene Expression by Coordinating with Polycomb Protein in Neurons

Bing Yao,[1,7,*] Yujing Li,[1,7] Zhiqin Wang,[1] Li Chen,[1,2,6] Mickael Poidevin,[1] Can Zhang,[3] Li Lin,[1] Feng Wang,[1] Han Bao,[1] Bin Jiao,[1] Junghwa Lim,[1] Ying Cheng,[1] Luoxiu Huang,[1] Brittany Lynn Phillips,[4] Tianlei Xu,[2] Ranhui Duan,[5] Kenneth H. Moberg,[3] Hao Wu,[2] and Peng Jin[1,8,*]

[1]Department of Human Genetics, Emory University School of Medicine, Atlanta, GA 30322, USA
[2]Department of Biostatistics and Bioinformatics, Emory University School of Medicine, Atlanta, GA 30322, USA
[3]Department of Cell Biology, Emory University School of Medicine, Atlanta, GA 30322, USA
[4]Department of Pharmacology, Emory University School of Medicine, Atlanta, GA 30322, USA
[5]State Key Laboratory of Medical Genetics, School of Life Science, Central South University, Changsha, Hunan 410078, China
[6]Present address: Department of Health Outcomes Research and Policy, Auburn University Harrison School of Pharmacy, Auburn, AL 36849, USA
[7]These authors contributed equally
[8]Lead Contact
*Correspondence: bing.yao@emory.edu (B.Y.), peng.jin@emory.edu (P.J.)
https://doi.org/10.1016/j.molcel.2018.07.005

SUMMARY

A ten-eleven translocation (TET) ortholog exists as a DNA N^6-methyladenine (6mA) demethylase (DMAD) in *Drosophila*. However, the molecular roles of 6mA and DMAD remain unexplored. Through genome-wide 6mA and transcriptome profiling in *Drosophila* brains and neuronal cells, we found that 6mA may epigenetically regulate a group of genes involved in neurodevelopment and neuronal functions. Mechanistically, DMAD interacts with the Trithorax-related complex protein Wds to maintain active transcription by dynamically demethylating intragenic 6mA. Accumulation of 6mA by depleting DMAD coordinates with Polycomb proteins and contributes to transcriptional repression of these genes. Our findings suggest that active 6mA demethylation by DMAD plays essential roles in fly CNS by orchestrating through added epigenetic mechanisms.

INTRODUCTION

Cytosine methylation at the 5-carbon position (5-methylcytosine; 5mC) is a critical repressive epigenetic mark in the mammalian genome (Bird, 2002; Ma et al., 2010; Schübeler, 2015; Spivakov and Fisher, 2007). 5mC is generally viewed as a stable and irreversible covalent modification to DNA; however, the fact that ten-eleven translocation (TET) proteins can oxidize 5mC to 5-hydroxymethylcytosine (5hmC) and downstream derivatives gives us a new perspective on the plasticity of 5mC-dependent regulatory processes (Zhang et al., 2012). Cytosine modifications exist in bacteria, archaea, viruses, fungi, vertebrates, and plants, but their presence and functions remain controversial in model organisms such as *Drosophila* (Lyko et al., 2000; Zhang et al., 2015b).

Surprisingly, a TET ortholog with unknown function exists in the fly genome (Dunwell et al., 2013). Recent studies demonstrated that this *Drosophila* TET ortholog could demethylate the DNA modification N^6-methyladenine (6mA) in eukaryotes (Zhang et al., 2015b), a prevalent DNA modification previously only found in bacteria. 6mA was also recently found in algae, worms, fungi, and mammals (Fu et al., 2015; Greer et al., 2015; Koziol et al., 2016; Mondo et al., 2017; Wu et al., 2016; Zhang et al., 2015b). Although this *Drosophila* TET ortholog was identified as a 6mA demethylase (DMAD) (Zhang et al., 2015b), the precise molecular functions of 6mA and DMAD in the *Drosophila* genome remain unknown.

RESULTS

DMAD Depletion in *Drosophila* Brain Results in Brain Developmental Defects Accompanied by 6mA Accumulation

Gene expression analyses across adult fly tissues show that DMAD is highly expressed in fly brains (Figure 1A). Previous work showed that DMAD is essential for development, since only a small fraction of DMAD null mutants can survive through the pupa stage, although these mutants die within 3 days post-eclosion (Zhang et al., 2015b). To understand the role DMAD plays in fly brain function, we generated multiple transgenic lines, P{*UASp-artimiR-DMAD*}, which carry artificial microRNAs (miRNAs) targeting DMAD. Consistent with DMAD null flies (Zhang et al., 2015b), ubiquitous expression of artificial miRNAs against DMAD using the driver *Tubulin-GAL4* (DMAD-knockdown [DMAD-KD]) resulted in developmental defects, as most flies died pre-eclosion (Figure S1A). The mRNA of DMAD was effectively depleted in two DMAD-KD lines (Figure S1B). Neuronal-specific DMAD-nKD under the control of the pan-neuronal driver

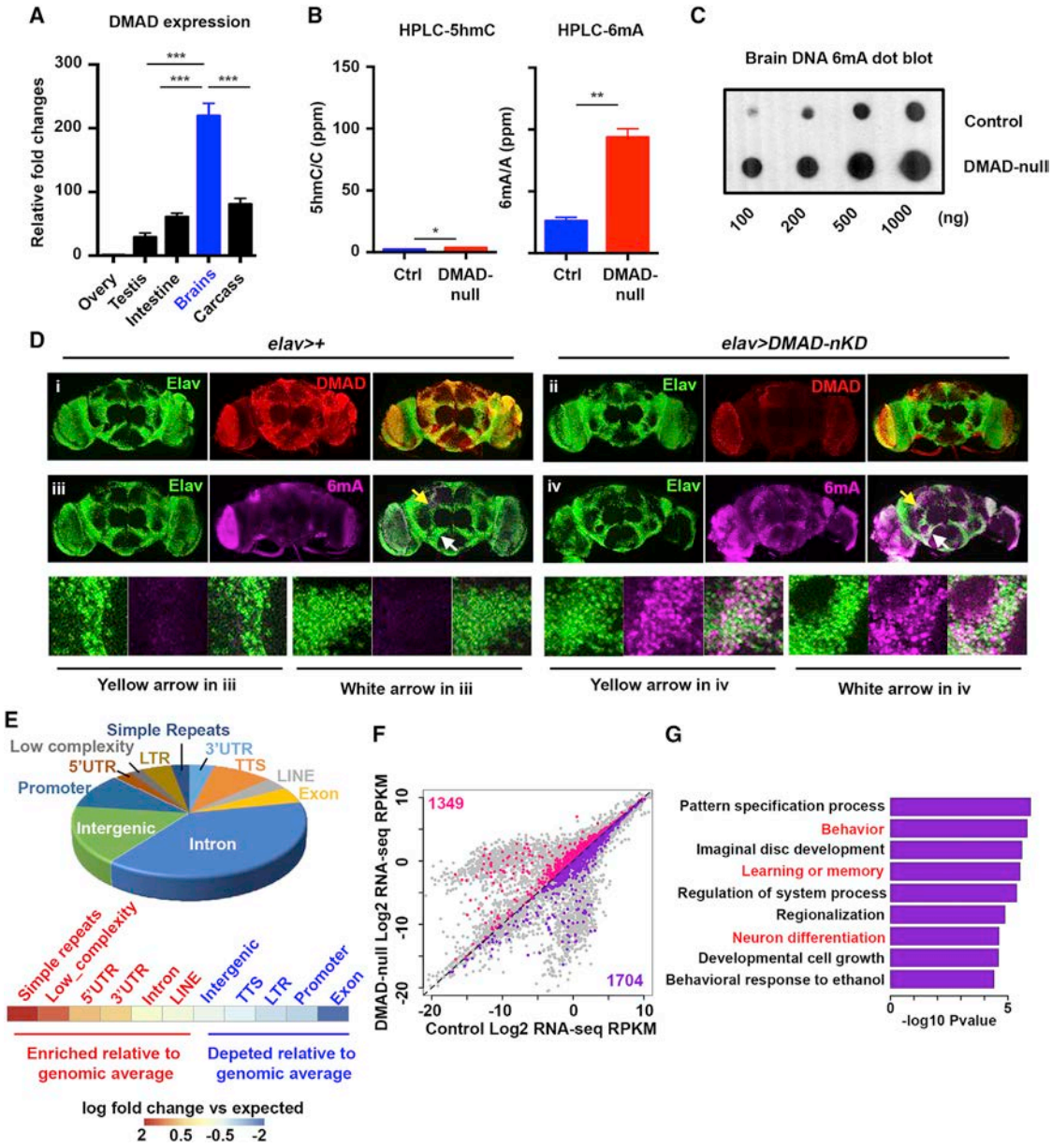

Figure 1. DMAD Demethylates Intragenic 6mA in *Drosophila* Brains

(A) qRT-PCR across fly tissues revealed high levels of DMAD expression in heads compared to other somatic tissues (n = 3). *p < 0.05; **p < 0.01; ***p < 0.001. Welch two sample t tests.

(B) High-resolution HPLC quantification of 5hmC and 6mA in *Drosophila* brains in the presence (Ctrl) and absence (DMAD null) of DMAD. 5hmC divided by total C and 6mA divided by total A are shown as percentages per million (ppm) nucleotides. There was a ~3.6-fold increase in 6mA with DMAD depletion, while 5hmC was undetectable (n = 2).

(C) Dot blots using an antibody specific for 6mA confirmed the accumulation of 6mA in DMAD null fly brains.

(D) Confocal images of adult brains stained with anti-Elav (green), anti-DMAD (red), and anti-6mA (purple) in the background of *elav-Gal4* alone (i and iii) or in combination with *UAS-DMAD* miRNA (DMAD-nKD) (ii and iv). Enlarged views of areas indicated with arrows in iii and iv are shown below. Compared to control brains (*elav-Gal4* alone), DMAD-KD significantly increased nuclear 6mA levels in *elav*-expressing cells.

(E) Genomic annotation of gain-of-6mA regions in DMAD null fly brains revealed their intragenic characteristics. A heatmap shows the enrichment of each genomic feature versus expected values.

(F) Plot of the global transcriptome in control and DMAD null fly brains obtained from RNA-seq (n = 2). Genes bearing gain-of-6mA regions in their gene bodies were highlighted (upregulated genes are in pink, and downregulated genes are in purple).

(G) Gene Ontology analysis was performed on a subset of downregulated genes in (F) (purple). Log_2 fold change, (KD: WT) < −0.5, was applied as the threshold cutoff. Several biological processes involved in neurodevelopment and neuronal functions were enriched and are highlighted in red.

Data are presented as mean ± SEM.

Elav-GAL4 resulted in mushroom body (MB) abnormalities in adult flies. MBs are one of the best characterized brain regions involved in learning and memory in flies and are composed of axon-like fiber structures forming Fas-II-positive neuronal lobes (the α-, β- and γ-lobes) (Figure S1C; Heisenberg, 2003). In DMAD-nKD flies, the β-lobes, which normally end at the midline cleft, often crossed the midline. Also, DMAD-nKD flies also bore accessory MB phenotypes, such as missing or misdirected α- and β-lobes, as well as truncated or overbranched lobes (Figure S1C). These data suggest that DMAD may contribute to neuronal morphology, development, and function in the fly brain.

DMAD possesses enzymatic activity as a DNA 6mA demethylase in fly ovaries (Zhang et al., 2015b). To confirm this enzymatic activity in fly brains, we applied highly sensitive ultra-high-performance liquid chromatography-tandem mass spectrometry (UHPLC-MS/MS) to quantify 6mA and 5hmC abundance in control and DMAD null fly brains (Schübeler, 2015; Zhang et al., 2015b). The DMAD protein level was abolished entirely in DMAD null fly brains (Figure S1D). Consistent with previous findings, a shallow level of 5hmC was detectable in control fly brains (Dunwell et al., 2013; Raddatz et al., 2013), and DMAD depletion had minimal effects on 5hmC abundance (average 5hmC/C [cytosine] percentage per million nucleotide [ppm] increased from 2.29 to 3.44). In contrast to 5hmC, we detected higher levels of 6mA in control fly brains (average, 26.0 6mA/A [adenosine] ppm). Importantly, 6mA levels significantly increased 4-fold in DMAD null flies relative to controls (Figure 1B; Table S1). We confirmed this finding with dot blots, using a 6mA-specific antibody (Figure 1C). To further confirm that DMAD is a 6mA demethylase in fly neuronal cells, we performed 6mA immunostaining in control and DMAD-nKD flies. Neuronal-specific DMAD-KD using *Elav-Gal4* resulted in 6mA accumulation explicitly in *Elav-expressing* neuronal cells (Figure 1D). Also, *in vitro*, 6mA demethylation assays using double-stranded synthetic oligonucleotide substrates indicated that recombinant DMAD demethylates 6mA (Figures S1E and S1F). These findings provide strong support for the role of DMAD as 6mA demethylase in fly brains. One recent study suggested that DMAD can generate 5-hydroxymethylcytosine in RNA molecules (5hmrC), and DMAD-deficient S2 cells showed decreased 5hmrC (Delatte et al., 2016). To test this, we applied UHPLC-MS/MS to precisely quantify the 5hmrC levels in control, DMAD null fly brains, and DMAD-KD S2 cells. Extremely low levels of 5hmrC (0.2 ppm in fly brains and 2 ppm in S2 cells versus 25–100 ppm 6mA) were detected (Figure S1G), indicating that 6mA is a major substrate for DMAD in the *Drosophila* genome. No significant change in RNA for 5mC was observed (Figure S1H). Furthermore, we tested the 6mA levels in an AlkB neuronal KD, which is a homolog of mammalian 6mA putative demethylase Alkbh1 (Wu et al., 2016). No substantial 6mA accumulation was found in the absence of AlkB (Figure S1I), suggesting that DMAD could be the key 6mA demethylase in *Drosophila*. Since the key residues responsible for DMAD's demethylation activities were defined (Zhang et al., 2015b), we generated the recombinant DMAD catalytic domain (DMAD-CD) and its catalytically dead mutations (DMAD-CD-mut) to test their direct demethylation activity *in vitro*. The DMAD-CD has comparable demethylation activities with bacterial 6mA demethylase ALKB. In agreement with previous observations, the DMAD-CD mutant displayed drastic and significant reduced enzymatic activity for 6mA demethylation (Figure S1J).

To understand the molecular functions of active 6mA demethylation, we generated genome-wide 6mA maps from control and DMAD null fly brains. We immunoprecipitated 6mA-containing DNA isolated from ~1,000 dissected fly brains and used high-throughput sequencing to identify genomic loci enriched for 6mA. The specificity and reproducibility of 6mA immunoprecipitation were ensured using competitive elution with excess 6mA. Overall, the number of binned genomic regions containing high 6mA reads was larger in DMAD null brains (Figure S1K). We identified 5,340 high-confidence gain-of-6mA regions in DMAD null mutants relative to control flies. Thus, these regions represent active 6mA demethylation loci in wild-type brains (Figure S1L; Table S1). Genomic annotation of gain-of-6mA regions revealed that a large percentage of 6mA increases occurred in intragenic regions, with enrichment in introns and UTRs compared to expected values. Consistent with previous observations (Zhang et al., 2015b), gain-of-6mA regions were also enriched for several classes of repetitive elements relative to expected values (Figure 1E).

To investigate the connection between dynamic 6mA demethylation and global transcriptome changes, we performed RNA sequencing (RNA-seq) on control and DMAD null fly brains (Table S1). 1,704 downregulated and 1,349 upregulated genes bearing significant accumulations of 6mA on their gene bodies upon DMAD depletion were found (Figure 1F, purple and pink, respectively). Gene Ontology (GO) analysis of these downregulated genes showed enrichment in several key ontology terms reflecting DMAD null fly phenotypes, such as behavior, learning, memory, and neuronal differentiation (Figure 1G). In contrast, upregulated genes were enriched in general development-related functions (Figure S1M). To further investigate the potential regulatory mechanisms underlying 6mA-mediated gene regulation, we searched for common motifs within gain-of-6mA regions and predicted possible binding factors based on these motifs. Intriguingly, the 6mA motif "AGAAGGAG" in fly brains was previously found in *C. elegans* (Greer et al., 2015), potentially suggesting a conserved mechanism of 6mA regulation across species. Several known DNA-binding proteins, such as transcription factors Aef1 and Adf1, were predicted to bind to these regions (Figure S1N). Interestingly, both Aef1 and Adf1 define Polycomb response elements (PREs) by interacting with Polycomb PRC1 component Pc (Orsi et al., 2014). These findings inspired us to explore the possible interplay between 6mA and histone modifiers.

6mA Is Associated with the Binding of Polycomb Protein in *Drosophila* Neuronal Cells

To gain in-depth mechanistic insight into 6mA dynamics and gene regulation, we assessed the effects of reduced DMAD expression in a fly neuronal cell line, BG3C2 (Ui et al., 1994), which was extensively analyzed in the modENCODE project. DMAD-KD was performed using double-stranded RNAs (dsRNAs), and successful DMAD-KD was confirmed by qPCR and western blots (Figures 2A and 2B). Consistent with our data in fly brains, DMAD-KD in the BG3C2 neuronal cell line also led to an overall increase in 6mA levels (Figure 2C). A total of 6,093 gain-of-6mA regions were identified in BG3C2 cells

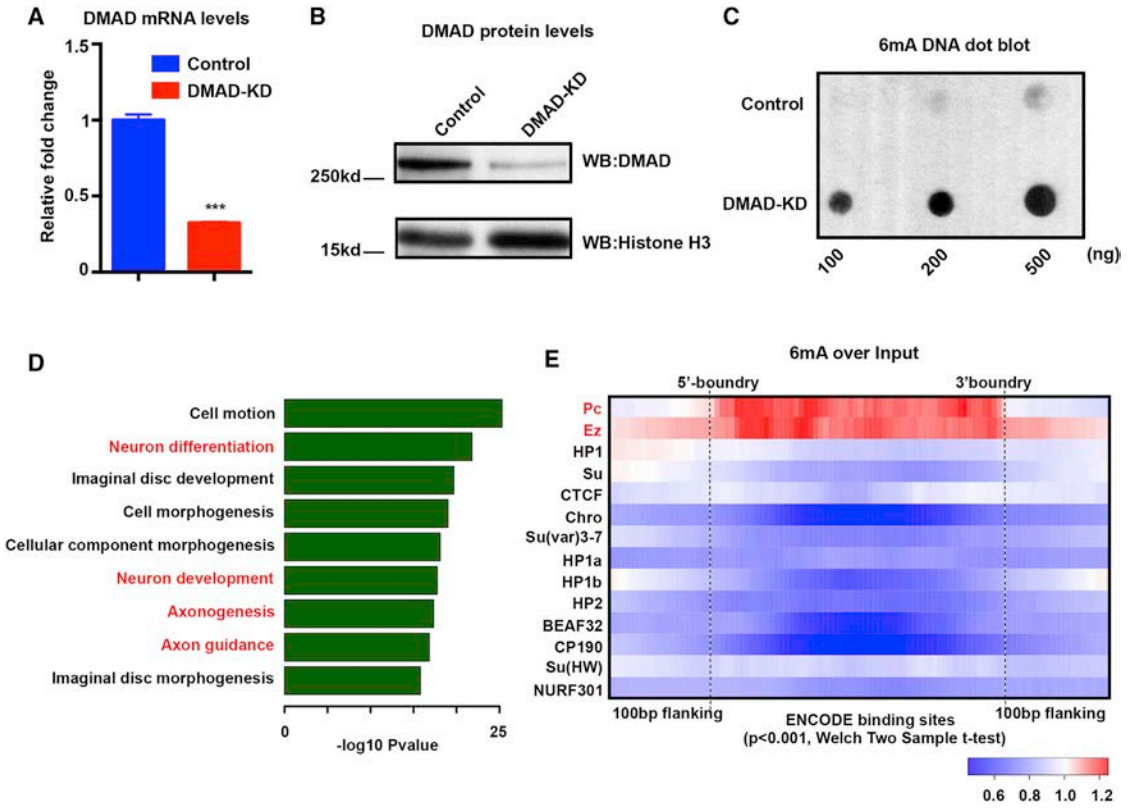

Figure 2. 6mA Is Enriched at Polycomb-Binding Sites in Neuronal Cells

(A) qRT-PCR validated a 70% reduction in DMAD mRNA levels after double-strand small interfering RNA (siRNA) knockdown in BG3C2 cells. ***$p < 0.001$. Welch two sample t tests.

(B) Western blot using a DMAD-specific antibody confirmed effective DMAD-KD in BG3C2 cells. kd, knockdown.

(C) Dot blots using a 6mA-specific antibody confirmed 6mA accumulation in the absence of DMAD in BG3C2 cells.

(D) GO analysis showed specific enrichment for neurodevelopment and neuronal functions from downregulated genes carrying intragenic BG3C2 gain-of-6mA regions. The red font highlights the key Gene Ontology terms related to neurodevelopment.

(E) Average fold change in 6mA mapped reads versus non-enriched input DNA was calculated for various binned ChIP-chip regions of epigenetic regulators available from the modENCODE database. Average fold change is plotted in Heatmap view. Red (fold change > 1) indicates enrichment over input, while blue (fold change < 1) indicates depletion. 6mA was explicitly enriched at Polycomb-protein-binding sites. Enrichment and depletion were significant with p value < 0.001, Welch two-sample t tests.

Data are presented as mean ± SEM.

from replicates of control and DMAD-KD cells (Figure S2A; Table S2). In parallel, RNA-seq analyses were performed for control and DMAD-KD BG3C2 cells (Table S2). Like our findings in the brain, gain-of-6mA regions were associated with intragenic regions such as introns and repetitive elements relative to expected values (Figure S2B). Also, GO analysis of downregulated genes containing intragenic gain-of-6mA regions showed preferential enrichment for neuronal functions (Figure 2D). Conversely, upregulated genes having gain-of-6mA regions were enriched for general developmental pathways (Figure S2C). Furthermore, we found 42.5% and 21.2% of neuronally expressed long interspersed elements (LINEs) and long terminal repeats (LTRs), respectively, overlapped with gain-of-6mA regions (Figure S2D), consistent with the previous report that 6mA could potentially impact transposon expression (Zhang et al., 2015b). To confirm the 6mA-immunoprecipitation sequencing (IP seq) data, we performed restriction enzyme digests to validate 6mA differential loci found by BG3C2 6mA-IP (STAR Methods). As expected, the qPCR results were consistent with our 6mA-IP findings that these loci contain methylation on the adenines and 6mA accumulated upon DMAD-KD (Figure S2E).

To directly link 6mA to epigenetic modulators, we used chromatin immunoprecipitation (ChIP)-chip data available for BG3C2 cells from modENCODE (Ho et al., 2014) to investigate 6mA enrichment at epigenetic regulator binding sites. We calculated 6mA reads over non-enriched input at the epigenetic regulator binding sites, finding that 6mA is explicitly enriched at Polycomb-protein-binding sites, including Polycomb complex components Pc, dRING, Psc, and Ez, but not other epigenetic regulators such as HP1, Su(var)3-7, and CTCF (Figures 2E and S2F). Moreover, DMAD-KD led to further 6mA accumulation, primarily on Polycomb-binding sites (Figure S2G). Therefore, 6mA could potentially work cooperatively with Polycomb proteins to mediate epigenetic regulation of gene expression.

DMAD Coordinates with the Trithorax-Related Protein Wds to Regulate Gene Expression by Maintaining 6mA Homeostasis at a Group of Neuronal Genes

To investigate the regulatory roles of DMAD in gene expression, we performed ChIP-seq to map genome-wide DMAD-binding sites in BG3C2 cells (Table S3). Genome-wide annotation of DMAD-binding sites revealed its general intragenic enrichment, such as exons (Figure S3A). Interestingly, 6mA accumulated on introns and UTRs upon DMAD-KD (Figure 3A), consistent with genomic annotations of gain-of-6mA regions (Figure 1E). We focused next on the DMAD-bound genes bearing accumulation of 6mA in their gene bodies in the DMAD-KD cells, as they could be modulated by DMAD through 6mA demethylation. These genes were further separated into downregulated and upregulated genes (Figures 3B and S3B). Interestingly, instead of more exon enrichment of general DMAD-binding sites, DMAD showed more enrichment in introns of these genes, coinciding with 6mA accumulation in introns (Figures 1E and S2B). These genes were subcategorized based on the DMAD intragenic binding regions (Figures 3C and S3C). In general, DMAD-binding sites and 6mA accumulation upon DMAD-KD were associated with the same genomic regions. For instance, a remarkable 95% and 98% of intronic 6mA accumulations were found in genes with DMAD binding to their introns, showing that DMAD influenced these genes through their intronic 6mA demethylation (Figures 3C and S3C, respectively). Interestingly, 76% of introns in downregulated genes harbored gain-of-6mA regions and DMAD-binding sites in the exact intronic regions, further suggesting that DMAD could demethylate 6mA *in cis* (Figure 3D). Similar observations were also found in the upregulated genes, although the percentage of DMAD and gain-of-6mA binding to the same introns showed a lesser extent (Figure S3D). Given that DMAD depletion downregulated a group of critical neuronal genes, we hypothesized that DMAD may coordinate with transcriptional activators to maintain active expression profiles. Using the modENCODE ChIP-chip dataset, we found that binding sites of Wds (Will die slowly), a protein component of the Trithorax-related complex that is responsible for gene activation (Mohan et al., 2011), substantially overlapped with DMAD-binding sites (Figure 3E). This overlap appeared specific, as the binding sites of many other transcription factors available in ChIP-chip datasets did not overlap with DMAD (Figure S3E). Wds is the *Drosophila* ortholog of the mammalian WDR5 and has been shown to modulate active transcription with other Trithorax components through histone modifications (Herz et al., 2012; Mohan et al., 2011). Substantial overlap between Wds- and DMAD-binding sites suggested that Wds and DMAD may form a complex *in vivo*. To test this, we performed co-immunoprecipitation (coIP) experiments and observed a biochemical association between DMAD and Wds in BG3C2 cells (Figure 3F). Altered genes in DMAD-KD BG3C2 cells significantly overlapped with differentially expressed genes due to the Wds KD (Figures 3G and S3F). Interestingly, KD of Wds did not affect the global 6mA level, suggesting that the DMAD target recognition and demethylation are independent of Wds (Figures S3G and S3H). CoIP experiments using Wds and DMAD deletion constructs were performed to map their interaction domains. WD40 domains of Wds were critical for interaction with DMAD, as the complete removal of WD40 domains abolished its binding to DMAD (Figure S3I). On the other hand, the C-terminal DMAD (amino acids [aa] 1,657–2,860) bearing the Tet-JBP catalytic domain, but not the DMAD N-terminal region (aa 1–1,657) with the CXXC DNA-binding domain, was responsible for association with Wds (Figure S3J). Additional fine mapping revealed that DMAD 1,657–2,666, including the 6mA catalytic domain (aa 1,796–2,666), bind to Wds (Figure S3K). These data together suggest that the DMAD interaction with Wds is coupled with its 6mA demethylation activities.

Since ChIP-chip assays, in general, have lower resolution than ChIP-seq, we performed Wds ChIP-seq in control and DMAD-KD BG3C2 cells (Table S3). 2,249 of 2,976 Wds peaks overlapped with DMAD-binding sites, actively supporting their cooperation in transcriptional control (Figure 3H). We further analyzed ChIP-seq data to understand the dynamic changes of Pc (Table S3) and Wds on DMAD-binding sites and gain-of-6mA regions. We computed the normalized Pc and Wds ChIP-seq read ratios at DMAD-binding sites using DMAD-KD over control cells. Wds, but not Pc, showed a significant reduction in the absence of DMAD (Figure 3I; blue indicates KD reads < WT reads; $p < 0.001$). Interestingly, substantial and significant increases in Pc binding were found on gain-of-6mA regions, suggesting potential crosstalk between 6mA and Polycomb proteins (Figure 3I; red indicates KD reads > WT reads; $p < 0.001$). Using micrococcal nuclease (MNase) digestion coupled with high-throughput sequencing, we investigated the dynamic changes of nucleosome positioning on intragenic DMAD-binding sites and gain-of-6mA regions when DMAD was depleted. DMAD-binding sites showed no significant changes in nucleosome occupancy when DMAD was KD (Figures S3L and S3N; $p = 0.4436$). In contrast, gain-of-6mA regions displayed significantly higher nucleosome occupancy in DMAD-KD cells (Figures S3M and S3O; $p < 0.0001$), possibly due to Polycomb recruitment for chromatin remodeling (Orsi et al., 2014). Our findings suggest that DMAD binds to specific sets of genes to modulate intragenic 6mA levels and coordinate with histone modifiers, thereby regulating gene expression.

6mA Dynamic Regulation by DMAD Coordinates with Trithorax- and Polycomb-Mediated Epigenetic Mechanisms

To further define DMAD-mediated epigenetic regulatory mechanisms, we focused on two sets of regions located in downregulated genes that have vital neuronal functions: (1) intragenic regions commonly bound by DMAD/Wds identified by ChIP-seq (Figure 4A, green) and (2) Gain-of-6mA regions in intragenic regions of genes identified in Figure 4A that are also commonly bound by DMAD/Wds (Figure 4B, pink). We examined the dynamic changes of Pc and Wds binding, as well as signature histone modifications, such as H3K27me3 and H3K4me3, at these loci in both control and DMAD-depleted cells. Intragenic regions bound by DMAD and/or Wds showed a slight but not significant reduction in Pc and H3K27me3 upon DMAD depletion. However, DMAD loss led to a substantial and significant reduction in Wds and the active histone marker H3K4me3 (Figure 4A). These data suggest that interplay between DMAD and Wds at these loci maintains transcriptional activation. We also noted that Wds

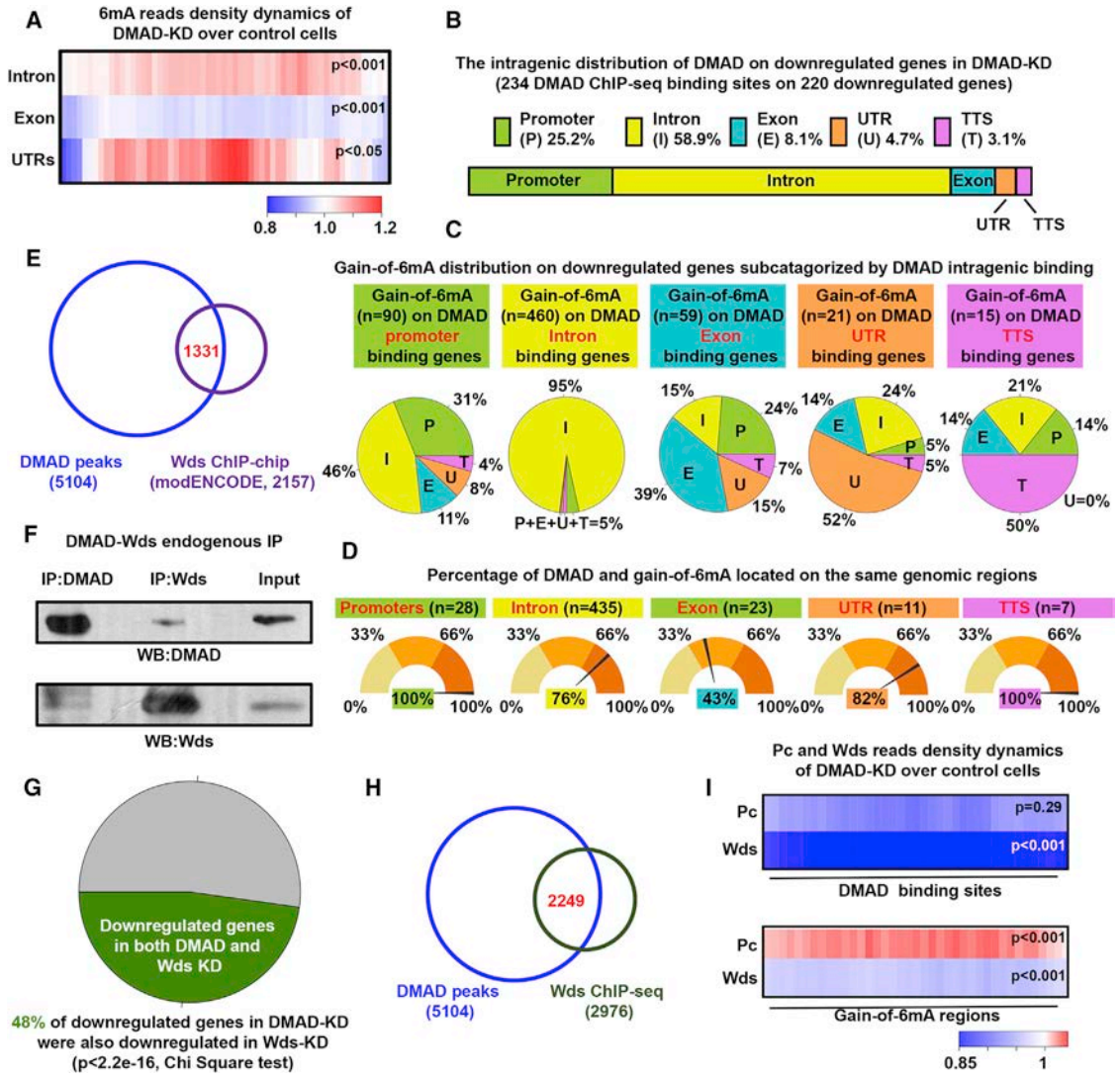

Figure 3. DMAD Demethylates Intragenic 6mA *In Cis* and Coordinates with Trithorax-Related Protein Wds to Regulate Gene Expression

(A) Average fold changes in 6mA reads between DMAD-KD and control were calculated for DMAD-occupied gene introns, exons, and UTRs. Enrichment or depletion of 6mA on these genomic regions were significant ($p < 0.001$ or $p < 0.05$, Welch two-sample t tests).

(B) Intragenic distributions of DMAD on DMAD-bound downregulated genes bearing accumulation of 6mA upon DMAD-KD was demonstrated proportionally. DMAD showed strong intronic enrichment on these genes. TTSs, transcription termination sites.

(C) The genes in (B) were further subcategorized based on the DMAD intragenic association. Gain-of-6mA distributions on each subset of genes were calculated, and the percentages are shown in the pie chart.

(D) The percentages of genes with both DMAD and gain-of-6mA binding to the same genomic feature were calculated.

(E) Venn diagram shows substantial and significant overlap between DMAD and Wds ChIP-chip data (binomial tests, $p < 0.001$).

(F) Co-immunoprecipitation experiments indicated a physical interaction between DMAD and Wds.

(G) Downregulated genes in the DMAD-KD cells showed significant overlapping with downregulated genes in the absence of Wds. Chi-square tests were performed. RNA-seq was performed in triplicates. Reads per kilobase per million mapped reads (RPKM) fold changes < -0.1 were included.

(H) Substantial overlap between DMAD and Wds ChIP-seq peaks suggested their functional coordination in regulating gene expression.

(I) Average fold change in Pc and Wds reads of DMAD-KD over control were calculated for both DMAD-binding sites and gain-of-6mA regions to explore Pc and Wds dynamics in these regions. The heatmap demonstrates a general decrease in both Pc and Wds at DMAD-binding sites when DMAD is depleted. A specific and significant increase in Pc binding on gain-of-6mA regions with DMAD depletion was found; Welch two-sample t tests, p values were indicated.

Data are presented as mean ± SEM.

was solely enriched at downregulated genes compared to non-enriched input (Figures S3P and S3Q). These data show that Wds, not Pc, is responsible for maintaining active expression of downregulated genes by interacting with DMAD. Genetic interaction between *DMAD* and *Wds* was further examined in the context of MB development. Either control (*elav* > +) or

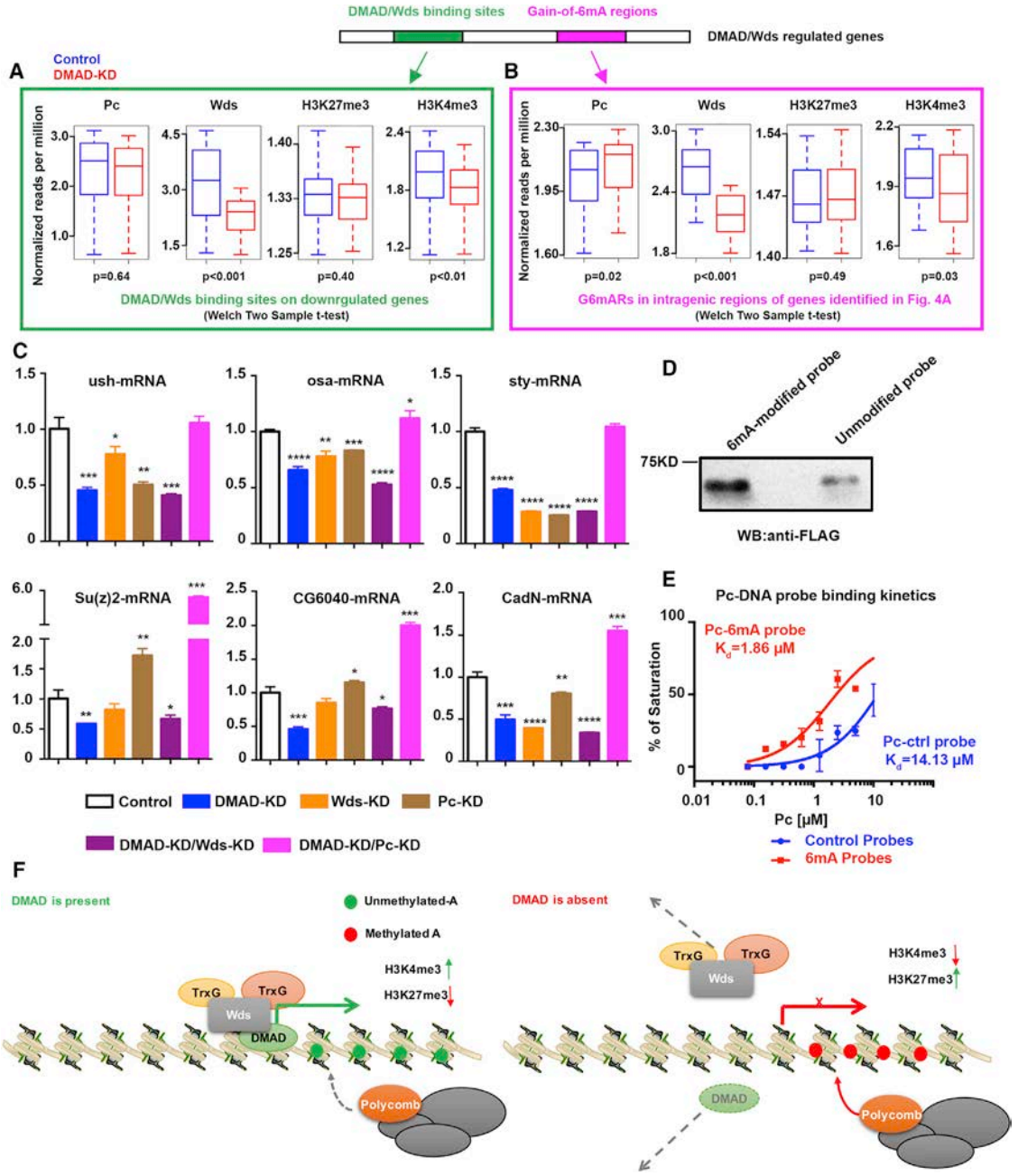

Figure 4. DMAD and 6mA Coordinate with Trithorax and Polycomb

(A) Average normalized reads (per million) dynamics in Pc, Wds, H3K27me3, and H3K4me3 at DMAD and Wds binding sites in genes downregulated in the absence of DMAD compared to the control are shown. Welch two-sample t tests; p values were indicated.

(B) Dynamics of average normalized reads (per million) in Pc, Wds, H3K27me3, and H3K4me3 at gain-of-6mA regions in intragenic regions of genes identified in (A) that were bound by DMAD and/or Wds. Welch two-sample t tests were performed; p values are indicated.

(C) Loci from (A) and (B) were further tested by qPCR for expression changes in the absence of DMAD, Wds, Pc, or combined depletion of DMAD and Pc, as well as DMAD and Wds. t tests were performed. *p < 0.05; **p < 0.01; ***p < 0.001; ****p < 0.0001.

(D) *In vitro* 6mA-Pc binding assays were performed to confirm direct correlation between 6mA and Pc.

(E) Pc binding kinetics to control and 6mA-modified probes showed that Pc displayed stronger binding to 6mA-modified DNA probes, as measured by fluorescence polarization assays.

(F) DMAD binds to a group of genes involved in neurodevelopment and neuronal functions. These genes are directly targeted by the Trithorax protein Wds to maintain an active transcription profile. Additionally, DMAD actively demethylates intragenic 6mA. In the absence of DMAD, Wds binding is reduced at these loci, and accumulation of intragenic 6mA recruits Polycomb proteins.

Data are presented as mean ± SEM.

pan-neuronal KD of Wds alone (*elav* > Wds^{RNAi}) does not affect α-lobe development. Pan-neuronal DMAD-KD (*elav* > $DMAD^{RNAi}$) causes an α-lobe defect that is significantly enhanced by the Wds KD (*elav* > Wds^{RNAi} + $DMAD^{RNAi}$) (Figure S3R). These results further confirmed functional interplay between DMAD and Wds and suggested that DMAD preceded and coordinated with Wds (Figure S3R). It also suggested that DMAD could partner with other Trithorax proteins in fly neurons. In agreement with this, Trithorax-related (Trr), a major H3K4 monomethyltransferase on enhancers, shared substantial overlap with DMAD (Herz et al., 2012) (Figure S3S).

Analysis of the second set of regions, which contains intragenic regions with increased 6mA upon DMAD-KD, shows that Pc binding was concomitantly and significantly enhanced (Figure 4B). This observation supports the notion that accumulation of intragenic 6mA due to DMAD loss could facilitate Polycomb recruitment to secure transcriptional repression at these loci. In the absence of DMAD, these data suggest a dynamic switch between Trithorax and Polycomb proteins at these downregulated genes.

We also investigated upregulated genes bearing DMAD-Wds binding and gain-of-6mA regions. The DMAD-binding sites on these genes displayed a more sophisticated modulating network, as indicated by the co-occupancy of Wds and Pc on these regions (Figures S3Q, S4A, and S4B). Importantly, GO analyses showed a specific enrichment of neurodevelopment and neuronal functional terms of these downregulated genes bound by DMAD, compared to general terms of upregulated genes (Figures S4C and S4D). This evidence implies that these downregulated genes could be the primary and direct targets of the DMAD-Wds complex related to neuronal development and functions.

To confirm the direct targets of DMAD and Wds, we independently KD DMAD, Wds, and Pc in BG3C2 cells. We also generated Pc-DMAD or Wds-DMAD double-KD cell lines (Figures S4E and S4F). Several loci shown in Figures 4A and 4B were tested for expression changes with these KD conditions. Similar genes were downregulated with either Wds KD or DMAD-KD, suggesting that Wds interacts and coordinates with DMAD to transcriptionally activate these genes. DMAD and Wds double KD showed synergistic effects on some loci. On the other hand, simultaneous depletion of DMAD and Pc sustained or stimulated gene expression, suggesting that Pc contributes to transcription repression (Figures 4C and S4F). However, we found that the KD of Pc itself resulted in a significant reduction of Wds (Figure S4E). The reduction in Wds might represent interesting crosstalk between TrxG and PcG (Schuettengruber et al., 2007). Considering this finding, we found that Pc KD resulted in reduced expression of some loci, possibly due to indirect effects of regulating Wds or other factor expression.

To obtain direct evidence that Pc preferentially binds to the 6mA mark, we synthesized DNA oligos with the gain-of-6mA and Pc common consensus sequence. The adenines in the consensus sequence were either methylated or unmodified as a negative control. Control and 6mA-containing probes were used for *in vitro* binding with recombinant FLAG-Pc produced from baculovirus. We found that Pc preferentially binds to 6mA-modified probes over controls *in vitro* (Figure 4D). By using a fluorescence-based DNA-binding assay (Hashimoto et al., 2014), we calculated the binding kinetics. FLAG-Pc bound to probes in a dose-dependent manner, and 6mA-modified probes showed stronger binding kinetics (K_d) to Pc relative to control (Figure 4E).

We generated three Pc deletion constructs covering the chromodomain (Pc aa 1–86), the linker region (Pc aa 75–228), and the cbox domain (Pc aa 222–390) and tested their binding affinities to control and 6mA-modified DNA oligos (Figures S4G and S4H) to define the 6mA-binding domain of Pc. Only the C-terminal Pc deletion (aa 222–390) bound to DNA probes and displayed a preference for 6mA-modified DNA probes, although with lower affinity than full-length Pc (Figure S4I). These data provide further evidence that 6mA has crosstalk with Polycomb proteins.

DISCUSSION

Cytosine modifications are rare, if present at all, in many model organisms, including worms (*C. elegans*) and insects (*D. melanogaster*). Speculation exists that 6mA could serve as a viable DNA modification in these organisms to epigenetically modulate transcription. Crosstalk between 6mA and H3K4me2 in *C. elegans* (Greer et al., 2015), as well as communication between cytosine methylation and H3K9me3 (Du et al., 2015), has been documented. Our data identify the molecular mechanisms of 6mA and its demethylase DMAD in gene regulation. Our findings suggest that 6mA plays epigenetic roles in regulating a group of genes involved in *Drosophila* neurodevelopment and neuronal functions. DMAD coordinates with the Trithorax-related protein Wds to maintain active transcription by removing intragenic 6mA. Depletion of DMAD results in a reduction of Wds and accumulation of 6mA. This 6mA accumulation recruits Polycomb Pc to implement transcriptional repression on these loci. Also, our data link DNA modification to histone modifications in insect cells (Figure 4F). It is worth noting that, like 5mC in mammals, the specific epigenetic role(s) of 6mA, either active or repressive, could depend on its co-factors and varies in different species, tissues, and cell types.

STAR★METHODS

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

- KEY RESOURCES TABLE
- CONTACT FOR REAGENT AND RESOURCE SHARING
- METHOD DETAILS
  - DMAD Null and DMAD-KD Fly Lines
  - Immunostaining
  - Cell Culture and RNAi
  - Isolation of Genomic DNA
  - UHPLC-MRM-MS/MS Analysis
  - Dot Blot
  - 6mA Immunoprecipitation
  - Chromatin Immunoprecipitation (ChIP)
  - Co-immunoprecipitation
  - Baculovirus Infection and FLAG-Tagged Pc Purification
  - Probe Labeling and Annealing
  - Pc and DNA Probe Binding Assays

SUPPLEMENTAL INFORMATION

Supplemental Information includes four figures and four tables and can be found with this article online at https://doi.org/10.1016/j.molcel.2018.07.005.

ACKNOWLEDGMENTS

We would like to thank S. Warren, K. Garber, and D. Cook for critical reading of the manuscript. This work was supported in part by the NIH (NS051630, NS079625, NS097206, NS091859, MH102690, AG052476, and HG008935 to P.J.), the March of Dimes (6-FY13-121 to P.J.), and the Emory Genetics Discovery Fund.

AUTHOR CONTRIBUTIONS

B.Y., Y.L., and P.J. conceived and designed the project. B.Y., Y.L., Z.W., M.P., C.Z., L.L., F.W., H.B., B.J., J.L., Y.C., and L.H. performed the experiments. B.Y., L.C., T.X., and H.W. performed the bioinformatics analyses. R.D., and K.H.M. contributed the reagents. B.L.P. edited the manuscript. B.Y. and P.J. wrote the manuscript. All authors commented on the manuscript.

DECLARATION OF INTERESTS

The authors declare no conflict of interest.

Received: December 18, 2017
Revised: June 5, 2018
Accepted: July 3, 2018
Published: August 2, 2018

REFERENCES

Ashburner, M., Ball, C.A., Blake, J.A., Botstein, D., Butler, H., Cherry, J.M., Davis, A.P., Dolinski, K., Dwight, S.S., Eppig, J.T., et al.; The Gene Ontology Consortium (2000). Gene Ontology: tool for the unification of biology. Nat. Genet. *25*, 25–29.

Bird, A. (2002). DNA methylation patterns and epigenetic memory. Genes Dev. *16*, 6–21.

Chen, K., Xi, Y., Pan, X., Li, Z., Kaestner, K., Tyler, J., Dent, S., He, X., and Li, W. (2013). DANPOS: dynamic analysis of nucleosome position and occupancy by sequencing. Genome Res. *23*, 341–351.

Delatte, B., Wang, F., Ngoc, L.V., Collignon, E., Bonvin, E., Deplus, R., Calonne, E., Hassabi, B., Putmans, P., Awe, S., et al. (2016). RNA biochemistry. Transcriptome-wide distribution and function of RNA hydroxymethylcytosine. Science *351*, 282–285.

Du, J., Johnson, L.M., Jacobsen, S.E., and Patel, D.J. (2015). DNA methylation pathways and their crosstalk with histone methylation. Nat. Rev. Mol. Cell Biol. *16*, 519–532.

Dunwell, T.L., McGuffin, L.J., Dunwell, J.M., and Pfeifer, G.P. (2013). The mysterious presence of a 5-methylcytosine oxidase in the *Drosophila* genome: possible explanations. Cell Cycle *12*, 3357–3365.

Feng, H., Conneely, K.N., and Wu, H. (2014). A Bayesian hierarchical model to detect differentially methylated loci from single nucleotide resolution sequencing data. Nucleic Acids Res. *42*, e69.

Fu, Y., Luo, G.Z., Chen, K., Deng, X., Yu, M., Han, D., Hao, Z., Liu, J., Lu, X., Dore, L.C., et al. (2015). N^6-methyldeoxyadenosine marks active transcription start sites in *Chlamydomonas*. Cell *161*, 879–892.

Greer, E.L., Blanco, M.A., Gu, L., Sendinc, E., Liu, J., Aristizábal-Corrales, D., Hsu, C.H., Aravind, L., He, C., and Shi, Y. (2015). DNA methylation on N^6-adenine in *C. elegans*. Cell *161*, 868–878.

Hashimoto, H., Olanrewaju, Y.O., Zheng, Y., Wilson, G.G., Zhang, X., and Cheng, X. (2014). Wilms tumor protein recognizes 5-carboxylcytosine within a specific DNA sequence. Genes Dev. *28*, 2304–2313.

Heinz, S., Benner, C., Spann, N., Bertolino, E., Lin, Y.C., Laslo, P., Cheng, J.X., Murre, C., Singh, H., and Glass, C.K. (2010). Simple combinations of lineage-determining transcription factors prime *cis*-regulatory elements required for macrophage and B cell identities. Mol. Cell *38*, 576–589.

Heisenberg, M. (2003). Mushroom body memoir: from maps to models. Nat. Rev. Neurosci. *4*, 266–275.

Herz, H.M., Mohan, M., Garruss, A.S., Liang, K., Takahashi, Y.H., Mickey, K., Voets, O., Verrijzer, C.P., and Shilatifard, A. (2012). Enhancer-associated H3K4 monomethylation by Trithorax-related, the *Drosophila* homolog of mammalian Mll3/Mll4. Genes Dev. *26*, 2604–2620.

Ho, J.W., Jung, Y.L., Liu, T., Alver, B.H., Lee, S., Ikegami, K., Sohn, K.A., Minoda, A., Tolstorukov, M.Y., Appert, A., et al. (2014). Comparative analysis of metazoan chromatin organization. Nature *512*, 449–452.

Huang da, W., Sherman, B.T., and Lempicki, R.A. (2009). Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. Nat. Protoc. *4*, 44–57.

Ji, H., Vokes, S.A., and Wong, W.H. (2006). A comparative analysis of genome-wide chromatin immunoprecipitation data for mammalian transcription factors. Nucleic Acids Res. *34*, e146.

Jin, Y., Tam, O.H., Paniagua, E., and Hammell, M. (2015). TEtranscripts: a package for including transposable elements in differential expression analysis of RNA-seq datasets. Bioinformatics *31*, 3593–3599.

Koziol, M.J., Bradshaw, C.R., Allen, G.E., Costa, A.S.H., Frezza, C., and Gurdon, J.B. (2016). Identification of methylated deoxyadenosines in vertebrates reveals diversity in DNA modifications. Nat. Struct. Mol. Biol. *23*, 24–30.

Langmead, B., Trapnell, C., Pop, M., and Salzberg, S.L. (2009). Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. Genome Biol. *10*, R25.

Luo, G.Z., Wang, F., Weng, X., Chen, K., Hao, Z., Yu, M., Deng, X., Liu, J., and He, C. (2016). Characterization of eukaryotic DNA N(6)-methyladenine by a highly sensitive restriction enzyme-assisted sequencing. Nat. Commun. *7*, 11301.

Lyko, F., Ramsahoye, B.H., and Jaenisch, R. (2000). DNA methylation in *Drosophila melanogaster*. Nature *408*, 538–540.

Ma, D.K., Marchetto, M.C., Guo, J.U., Ming, G.L., Gage, F.H., and Song, H. (2010). Epigenetic choreographers of neurogenesis in the adult mammalian brain. Nat. Neurosci. *13*, 1338–1344.

Mohan, M., Herz, H.M., Smith, E.R., Zhang, Y., Jackson, J., Washburn, M.P., Florens, L., Eissenberg, J.C., and Shilatifard, A. (2011). The COMPASS family of H3K4 methylases in *Drosophila*. Mol. Cell. Biol. *31*, 4310–4318.

Mondo, S.J., Dannebaum, R.O., Kuo, R.C., Louie, K.B., Bewick, A.J., LaButti, K., Haridas, S., Kuo, A., Salamov, A., Ahrendt, S.R., et al. (2017). Widespread adenine N6-methylation of active genes in fungi. Nat. Genet. *49*, 964–968.

Orsi, G.A., Kasinathan, S., Hughes, K.T., Saminadin-Peter, S., Henikoff, S., and Ahmad, K. (2014). High-resolution mapping defines the cooperative architecture of Polycomb response elements. Genome Res. *24*, 809–820.

Quinlan, A.R., and Hall, I.M. (2010). BEDTools: a flexible suite of utilities for comparing genomic features. Bioinformatics *26*, 841–842.

Raddatz, G., Guzzardo, P.M., Olova, N., Fantappié, M.R., Rampp, M., Schaefer, M., Reik, W., Hannon, G.J., and Lyko, F. (2013). Dnmt2-dependent methylomes lack defined DNA methylation patterns. Proc. Natl. Acad. Sci. USA *110*, 8627–8631.

Schübeler, D. (2015). Function and information content of DNA methylation. Nature *517*, 321–326.

Schuettengruber, B., Chourrout, D., Vervoort, M., Leblanc, B., and Cavalli, G. (2007). Genome regulation by polycomb and trithorax proteins. Cell *128*, 735–745.

Shen, L., Shao, N., Liu, X., and Nestler, E. (2014). ngs.plot: quick mining and visualization of next-generation sequencing data by integrating genomic databases. BMC Genomics *15*, 284.

Spivakov, M., and Fisher, A.G. (2007). Epigenetic signatures of stem-cell identity. Nat. Rev. Genet. *8*, 263–271.

Szulwach, K.E., Li, X., Li, Y., Song, C.X., Wu, H., Dai, Q., Irier, H., Upadhyay, A.K., Gearing, M., Levey, A.I., et al. (2011). 5-hmC-mediated epigenetic dynamics during postnatal neurodevelopment and aging. Nat. Neurosci. *14*, 1607–1616.

Tie, F., Banerjee, R., Fu, C., Stratton, C.A., Fang, M., and Harte, P.J. (2016). Polycomb inhibits histone acetylation by CBP by binding directly to its catalytic domain. Proc. Natl. Acad. Sci. USA *113*, E744–E753.

Trapnell, C., Roberts, A., Goff, L., Pertea, G., Kim, D., Kelley, D.R., Pimentel, H., Salzberg, S.L., Rinn, J.L., and Pachter, L. (2012). Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks. Nat. Protoc. *7*, 562–578.

Ui, K., Nishihara, S., Sakuma, M., Togashi, S., Ueda, R., Miyata, Y., and Miyake, T. (1994). Newly established cell lines from Drosophila larval CNS express neural specific characteristics. In Vitro Cell. Dev. Biol. Anim. *30A*, 209–216.

Wu, H., Wang, C., and Wu, Z. (2013). A new shrinkage estimator for dispersion improves differential expression detection in RNA-seq data. Biostatistics *14*, 232–243.

Wu, T.P., Wang, T., Seetin, M.G., Lai, Y., Zhu, S., Lin, K., Liu, Y., Byrum, S.D., Mackintosh, S.G., Zhong, M., et al. (2016). DNA methylation on N(6)-adenine in mammalian embryonic stem cells. Nature *532*, 329–333.

Yao, B., Lin, L., Street, R.C., Zalewski, Z.A., Galloway, J.N., Wu, H., Nelson, D.L., and Jin, P. (2014). Genome-wide alteration of 5-hydroxymethylcytosine in a mouse model of fragile X-associated tremor/ataxia syndrome. Hum. Mol. Genet. *23*, 1095–1107.

Zhang, Y., Liu, T., Meyer, C.A., Eeckhoute, J., Johnson, D.S., Bernstein, B.E., Nusbaum, C., Myers, R.M., Brown, M., Li, W., and Liu, X.S. (2008). Model-based analysis of ChIP-seq (MACS). Genome Biol. *9*, R137.

Zhang, L., Lu, X., Lu, J., Liang, H., Dai, Q., Xu, G.L., Luo, C., Jiang, H., and He, C. (2012). Thymine DNA glycosylase specifically recognizes 5-carboxylcytosine-modified DNA. Nat. Chem. Biol. *8*, 328–330.

Zhang, C., Robinson, B.S., Xu, W., Yang, L., Yao, B., Zhao, H., Byun, P.K., Jin, P., Veraksa, A., and Moberg, K.H. (2015a). The ecdysone receptor coactivator Taiman links Yorkie to transcriptional control of germline stem cell factors in somatic tissue. Dev. Cell *34*, 168–180.

Zhang, G., Huang, H., Liu, D., Cheng, Y., Liu, X., Zhang, W., Yin, R., Zhang, D., Zhang, P., Liu, J., et al. (2015b). N6-methyladenine DNA modification in *Drosophila*. Cell *161*, 893–906.

STAR★METHODS

KEY RESOURCES TABLE

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Antibodies		
Anti-N6-methyladenosine antibody	Synaptic Systems	202003; RRID: AB_2279214
Anti-H3K4me3	Abcam	ab8580; RRID: AB_306649
Anti-H3K27me3	Abcam	ab6002; RRID: AB_305237
Anti-Pc antibody (dN-19)	Santa Cruz	sc-15814; RRID: AB_672372
Anti-Wds	Novus Biologicals	40630002; RRID: AB_11030435
Anti-DMAD	Chen Lab	Zhang et al., 2015b
Anti-Fasciclin II	DSHB	1D4; RRID: AB_528235
Anti-HA (16B12)	Covance	MMS-101P; RRID: AB_2314672
Anti-Myc (9E10)	Thermo Fisher Scientific	MA1-980; RRID: AB_558470
Rabbit TrueBlot anti-Rabbit IgG HRP	Rockland	18-8816-33; RRID: AB_2610848
Mouse TrueBlot anti-Mouse IgG HRP	Rockland	18-8817-30; RRID: AB_2610849
Alexa Fluor 488 Anti-Rabbit IgG	Jackson ImmunoResearch	711-545-152; RRID: AB_2313584
Cy3 Anti-Rabbit IgG (H+L)	Jackson ImmunoResearch	711-165-152; RRID: AB_2307443
Anti-FLAG M2 Magnetic Beads	Sigma	M8823; RRID: AB_2637089
Anti-HA−Agarose	Sigma	A2095; RRID: AB_257974
EZview Red Anti-c-Myc Affinity Gel	Sigma	E6654; RRID: AB_10093201
Dynabeads Protein G	Thermo Fisher Scientific	10003D
Chemicals, Peptides, and Recombinant Proteins		
Recombinant Pc full length	This study	Pc full length
Recombinant Pc (aa 1-86)	This study	Pc (aa 1-86)
Recombinant Pc (aa 75-228)	This study	Pc (aa 75-228)
Recombinant Pc (aa 222-390)	This study	Pc (aa 222-390)
DMAD catalytic domain	This study	N/A
Agencourt AMPure XP beads	Beckman Coulter	A63880
Dynabeads MyOne Streptavidin C1	Thermo Fisher Scientific	65001
Cellfectin II Reagent	Thermo Fisher Scientific	10362100
Dpn I	New England Biolabs	R0176S
Calf intestinal phosphatase	New England Biolabs	M0290S
Phosphodiesterase I from Crotalus adamanteus venom	Sigma	P3243
Insulin solution human	Sigma	I9278
Sodium Ascorbate	Sigma	A4034
Alpha-ketoglutaric acid	Sigma	K3752
Lithium chloride	Sigma	L4408
Bacterial and Virus Strains		
E. coli (DH5a)-One Shot Top10	Thermo Fisher Scientific	C404003
Baculovirus	LakePharma	N/A
Critical Commercial Assays		
Qubit dsDNA HS Assay Kit	Thermo Fisher Scientific	Q32854
High Sensitivity DNA Analysis Kits	Agilent	5067-4627
NEBNext DNA Library Prep Reagent Set	New England Biolabs	E6000
HiScribe T7 High Yield RNA Synthesis Kit	New England Biolabs	E2040S
TruSeq RNA Sample Prep Kit V2	Illumina	RS-122-2001
Pierce BCA Protein Assay Kit	Thermo Fisher Scientific	23225

(*Continued on next page*)

Continued

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Q5 Site-Directed Mutagenesis Kit	New England Biolabs	E0554S
Pierce Biotin 3′ End DNA Labeling Kit	Thermo Fisher Scientific	89818
Deposited Data		
All sequencing data	This paper	GEO: GSE67855
Raw images	Mendeley Data	https://doi.org/10.17632/wy3jshg29m.3
Experimental Models: Cell lines		
BG3C2 cells	The Drosophila Genomics Resource Center (DGRC)	Ui et al., 1994
SF9 cells	Lab of Xiaodong Cheng	N/A
Experimental Models: Organisms/Strains		
DMAD[1]	Chen Lab	Zhang et al., 2015b
DMAD[2]	Chen Lab	Zhang et al., 2015b
DMAD RNAi	This study	N/A
Wds RNAi	Bloomington stock center	60399
Recombinant DNA		
pET- Pc-1-390	Harte Lab	Tie et al., 2016
pET- Pc-1-86	Harte Lab	Tie et al., 2016
pET- Pc-75-390	Harte Lab	Tie et al., 2016
pET- Pc-75-228	Harte Lab	Tie et al., 2016
pFastBac FPC	Addgene	Plasmid #1927
pFastBac-Flag-Pc-1-86	This study	N/A
pFastBac-Flag-Pc-75-228	This study	N/A
pFastBac-Flag-Pc-222-390	This study	N/A
pAc5.1-HA-Wds full length	This study	N/A
pAc5.1-HA-Wds-1-61	This study	N/A
pAc5.1-HA-Wds-1-187	This study	N/A
pAc5.1-HA-Wds-1-262	This study	N/A
pAc5.1-myc- DMAD full length	This study	N/A
pAc5.1-myc- DMAD-1-1657	This study	N/A
pAc5.1-myc-DMAD-1657-2860	This study	N/A
pAc5.1-myc-1658-1796	This study	N/A
pAc5.1-myc-1797-2666	This study	N/A
pAc5.1-myc-2667-2860	This study	N/A
Oligonucleotides		
artmiR-*DMAD*-1 s: 5′-ctagcagtCGATGTACTAGAATGGCTGG Atagttatattcaagcata TGCAGCCATTGTAGTACATCGgcg-3′	IDT	N/A
artmiR-*DMAD*-1-as: 5′-aattcgcCGATGTACTACAATGGCTGC Atatgcttgaatataacta TCCAGCCATTCTAGTACATCGactg-3′	IDT	N/A
artmiR-*DMAD*-2 s: 5′-ctagcagtCGCCTATGATCCCTATCAGT Atagttatattcaagcata TTCTGATAGGCATCATAGGCGgcg-3′	IDT	N/A
artmiR-*DMAD*-2-as: *5′-aattcgcCGCCTATGATGCCTATCAGAA tatgcttgaatataacta CTACTGATAGGGATCATAGGCGactg-3′*	IDT	N/A
6mA-modified probe for Pc binding assay: GAT CGA TCG ACA /iN6Me-dA/CA /iN6Me-dA/CA /iN6Me-dA/CA /iN6Me-dA/ CA /iN6Me-dA/CA /iN6Me-dA/CA /iN6Me-dA/CA /iN6Me-dA/GA TCG ATC GA	IDT	N/A
Reverse complemented probe: TCG ATC GAT CTG TGT GTG TGT GTG TGT CGA TCG ATC	IDT	N/A

(*Continued on next page*)

Continued

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Software and Algorithms		
Bowtie v.1.1.2	Langmead et al., 2009	N/A
MACS 1.4.2	Zhang et al., 2008	N/A
Ngsplot 2.61	Shen et al., 2014	N/A
HOMER v.4.9.1	Heinz et al., 2010	N/A
TopHat v.2.0.13	Trapnell et al., 2012	N/A
Cuffdiff v.2.2.1	Trapnell et al., 2012	N/A
Bedtools v.2.17.0	Quinlan and Hall, 2010	N/A
R/Bioconductor package DSS 3.7	Wu et al., 2013	N/A
CisGenome v.2.0	Ji et al., 2006	N/A
TEToolkit v.1.5.1	Jin et al., 2015	N/A
Dynamic analysis of nucleosome position and occupancy by sequencing (DANPOS) v.2	Chen et al., 2013	N/A

CONTACT FOR REAGENT AND RESOURCE SHARING

Further information and requests for resources and classifiers should be directed to and will be fulfilled by Bing Yao (bing.yao@emory.edu) or Peng Jin (peng.jin@emory.edu). There are no restrictions for use of the materials disclosed.

METHOD DETAILS

DMAD Null and DMAD-KD Fly Lines

DMAD null flies were described previously (Zhang et al., 2015b). DMAD transgenic RNAi lines, P{*uasp-2* × *artmiR-DMAD*}, in which two sets of 71 nt oligos containing a small hairpin sequence targeting the DMAD coding region under the control of the uasp promoter, were generated according to the method described previously (Zhang et al., 2015b). The sequences of these oligos are below.

artmiR-*DMAD*-1 s*: 5′*-ctagcagtCGATGTACTAGAATGGCTGGAtagttatattcaagcataTGCAGCCATTGTAGTACATCGgcg-3′
artmiR-*DMAD*-1-as*:*
5′-aattcgcCGATGTACTACAATGGCTGCAtatgcttgaatataactaTCCAGCCATTCTAGTACATCGactg-3′
artmiR-*DMAD*-2 s*:*
5′-ctagcagtCGCCTATGATCCCTATCAGTAtagttatattcaagcataTTCTGATAGGCATCATAGGCGgcg-3′
artmiR-*DMAD*-2-as*:*
5′-aattcgcCGCCTATGATGCCTATCAGAAtatgcttgaatataactaTACTGATAGGGATCATAGGCGactg-3′

Immunostaining

Fly brains were dissected in calcium-free 1x PBS and fixed in 4% paraformaldehyde for 1 hr on ice. Fixed tissues were permeabilized with 0.3% PTX (1x PBS + 0.3% Triton X-100) at room temperature for 30 min and then blocked in 0.1% PTX (1x PBS + 0.1% Triton X-100) + 5% Normal Goat Serum for 30 more min. Tissues were then incubated with anti-DMAD (1:200, Rb) or anti-6mA (1:500, Rb) in PBTG (PBST + 5% normal goat serum) overnight at 4°C. Anti-Fasciclin II monoclonal antibody (DHSB, 1D4) was used to stain the *Drosophila* mushroom body. The next day, tissues were washed 3 times with 1x PBST for 5 min and incubated with secondary antibody for 1 hr at room temperature. The tissues were washed 3 times with 1x PBST and mounted onto slides with Vectashield to prevent photobleaching. Samples were kept at −20°C until observed under confocal microscopy.

Cell Culture and RNAi

BG3C2 cells were cultured as described previously (Ui et al., 1994). For double stranded RNA (dsRNA) KD, a 580bp dsRNA fragment targeting DMAD coding region was generated via T7 promoter-driven *in vitro* transcription using an RT-PCR product as template. The following primers were used for RT-PCR: forward −5′gaaatCTCGAGtaatacgactcactatagggCGGAGCCAGTAGTTTTCAGC3′and reverse - 5′gaaatGAATTCtaatacgactcactatagggCATGGGGTTGATCTTCTCGT3′. The *in vitro* transcription was performed using HiScribe T7 High Yield RNA Synthesis Kit from NEB following manufacturer's instructions. dsRNA products were purified by two-step extractions with phenol: chloroform: isoamyl alcohol = 25:24:1, followed by chloroform alone. The extracted dsRNA was precipitated with absolute ethanol. The dsRNA was incubated with BG3C2 cells that had been grown in Shields and Sang M3 insect medium under FBS starvation for at least 24 hr. After 24-hour incubation with dsRNA, FBS was added to the culture media to a final

concentration of 3%. Fresh media containing dsRNA was replaced every 48 hr and the total incubation time was 144 hr to ensure effective KD. *Drosophila* S2 cells were cultured as described previously for co-immunoprecipitation experiments (Zhang et al., 2015a).

Isolation of Genomic DNA

Lysis buffer containing 100mM Tris-HCl (pH 8.5), 5mM EDTA (pH 8.0), 0.2% SDS, 200mM NaCl, and 20-25 μL of protease K (20mg/ml) was added to homogenized tissues or cells, mixed well, and incubated at 55°C overnight. After the overnight digestion, the lysates were brought to room temperature, and incubated with 5 μL of RNase A solution (20mg/ml) for at least 1 hr at room temperature. DNA was extracted by adding equal volume of phenol: chloroform: isoamyl alcohol at a ratio of 25:24:1 and centrifuged at 13,000 rpm for 10 min. Supernatant was transferred to clean tubes. Equal volume of isopropanol was then added to the supernatant and mixed well at room temperature to precipitate DNA. Once flocky DNA was visible, the precipitate was transferred to a new tube containing 1ml of 70% ethanol to wash. DNA was then collected by centrifugation and air-dried.

UHPLC-MRM-MS/MS Analysis

Genomic DNA was enzymatically digested into single nucleosides with a mixture of DNaseI, calf intestinal phosphatase, and snake venom phosphodiesterase I at 37°C for 12 h. After the enzymes were removed by ultrafiltration, the digested DNA was subjected to UHPLC-MS/MS analysis. HPLC fractionation of *Drosophila* m6dA and UHPLC-QTOF-MS/MS analysis were performed as described previously (Zhang et al., 2015b).

Dot Blot

Dot blot was performed as described previously (Szulwach et al., 2011) using 6mA rabbit polyclonal antibody. DNA samples were subjected to extensive RNase treatments before loading on Nitrocellulose membrane.

6mA Immunoprecipitation

Genomic DNA was sonicated to 200-300 bp fragments for 6mA enrichment using rabbit polyclonal antibody (Synaptic Systems) at 1:100 in 1x IP buffer containing 100mM Tris-HCl (pH7.4), 150mM NaCl and 0.05% Triton X-100. The DNA-antibody incubation was conducted on a rotating platform at 4°C overnight. Dynabeads Protein G (Novex by Life Technologies, REF 10009D 30mg/ml) were added the next day for additional 2 hr at 4°C. The beads were then washed six times at room temperature with 1x ice-cold IP buffer. After the wash, the immunoprecipitated DNA fragments were eluted by adding IP buffer containing 2.6mM 6mA for competitive elution. The elutions were repeated for three times at room temperature. The eluted DNA fragments were precipitated by isopropanol.

Chromatin Immunoprecipitation (ChIP)

BG3C2 cells (5-10 × 10^6) in the presence or absence of DMAD were fixed in 1% formaldehyde for 10 min at room temperature with gentle shaking, then 0.125 M final concentration of Glycine was added for additional 5-minute incubation to stop the fixation. Fixed cells were lysed on ice for 10 min in a NP-40 lysis buffer (10 mM HEPES/pH7.9, 0.5% NP-40, 1.5 mM $MgCl_2$, 10 mM KCl, 0.5 mM DTT and protease inhibitor cocktail) to release nuclei. After centrifugation at 4000 rpm for 5 min, the nuclear pellets were further lysed by sonication on ice in a nuclear lysis buffer (20 mM HEPES/pH7.9, 25% glycerol, 0.5% NP-40, 0.5% Triton X-100, 0.42 M NaCl, 1.5 mM $MgCl_2$, 0.2 mM EDTA and protease inhibitor cocktail), then centrifuged at 13,000 rpm for 10 min at 4°C. The supernatant was diluted with 2 volumes of dilution buffer (0.01% SDS, 1% Triton X-100, 1.2 mM EDTA, 167 mM NaCl, 16.7 mM Tris-HCl/pH8.0 and protease inhibitor cocktail). Immunoprecipitation was performed with desired antibodies for 6 hr to overnight at 4°C. After antibody incubation, 20 μL salmon sperm blocked DNA/protein G agarose (Upstate) were added and incubated for additional 1 hr. Precipitates were sequentially washed with TSE I (0.1% SDS, 1% Triton X-100, 2 mM EDTA, 150 mM NaCl, 20 mM Tris-HCl/pH 8.0), TSE II (0.1% SDS, 1% Triton X-100, 2 mM EDTA, 500 mM NaCl, 20 mM Tris-HCl/pH 8.0), TSE III (0.25 M LiCl, 1% NP-40, 1% deoxycholate, 1 mM EDTA, 10 mM Tris-HCl/pH 8.0), and then washed twice with TE buffer. The ChIP'ed DNA was eluted by elution buffer containing 1% SDS and 0.1 M $NaHCO_3$. NaCl at final concentration of 0.2M was added to the elution along with 1 μg RNaseA for reverse-crosslinking at 65°C for at least 6 hr. DNA fragments were purified using the PCR Purification Kit (QIAGEN). ChIP'ed DNA was subjected to library preparation.

Co-immunoprecipitation

BG3C2 or S2 cells were lysed for 10 min on ice with lysis buffer containing 25 mM Tris-HCl (pH7.5), 2mM $MgCl_2$, 5mM DTT, 0.5% Triton X-100, and 300mM NaCl in the presence of DNase and RNase. The cell lysate was sonicated before centrifugation at maximum speed for 10 min; the supernatant was incubated with antibodies overnight at 4°C. After antibody incubation, 20 μL salmon sperm blocked DNA/protein G agarose beads (Upstate) were added and incubated for additional 1 hr. Beads were then washed 3 times by lysis buffer. The immunoprecipitated proteins were eluted with 2x Laemmli Sample Buffer (Bio-Rad) followed by western blots.

Baculovirus Infection and FLAG-Tagged Pc Purification

SF9 cells were a gift from Dr. Xiaodong Cheng's lab in the Department of Biochemistry at the Emory University School of Medicine. SF9 cells were cultured in Sf-900 III SFM media from Thermo Fisher Scientific, and infected with Baculovirus generated with

pFastBac-FPC (full-length Pc) with FLAG tag obtained from Addgene. Baculovirus were produced by LakePharma. 72-hour post infection, SF9 cells were collected and lysed in lysis buffer containing 25mM Tris-Cl (pH 7.5), 300mM NaCl, 0.5% Triton X-100, and protease inhibitor tablet (Roche) for 30 min on ice followed by sonication at 1.5V, 0-3W for 20 s. Samples with 300 μL volume in each tube were sonicated 4 times. Lysate was centrifuged at 13,000 rpm for 10 min and supernatant was transferred to a new tube for IP purification. The supernatant was incubated with FLAG-M2 dynabeads (Sigma) and rotated at 4°C for 3 hr. Beads were then washed 5 times with 1x lysis buffer. To elute the FLAG-Pc, beads were incubated with the elution buffer containing 3X FLAG peptide at 100ug/ml for 2 hr at 4°C with gentle rotation. FLAG-Pc were tested by western blots.

Probe Labeling and Annealing

To compare Pc binding efficiency to 6mA DNA and unmodified DNA, two oligos were designed based on gain-of-6mA regions and the consensus sequence recognized by Pc (−5′-GAT CGA TCG A-CA CAC ACA CAC ACA CA-G ATC GAT CGA-3′).

The forward and reverse oligos were separately labeled using the Biotin 3′ End DNA Labeling Kit (Catalog number: 89818) from PierceFisher Scientific per manufacturer's instruction. Labeled oligos were precipitated with 0.1 volume of 3M NaAc (pH 5.2) and 3 volumes of absolute ethanol. Precipitated oligos were dissolved in 50 μL of nuclease free water and annealed into double stranded oligos.

Pc and DNA Probe Binding Assays

dsDNA probes with either unmodified A or 6mA were independently mobilized on Dynabeads Streptavidin MyOne C1 (Life Technologies) with 1x W/B buffer containing 25mM Tris-Cl (pH7.5), 1mM EDTA, and 1M NaCl at room temperature for 30 min with gentle rotation. The beads were then immobilized with probes and blocked with Blocker BSA (10x), Thermo Fisher Scientific (Catalog number: 37520) in TBS for 30 min at room temperature. Beads were washed twice with 1x lysis buffer. A range of purified FLAG-Pc protein concentrations (0.01 μM to 1 μM) was added to blocked beads and incubated for 1 hr at room temperature with gentle rotation. The beads were then washed extensively with 1x lysis buffer 5 times. FLAG-Pc was eluted in 2x Laemmli Sample Buffer (Bio-Rad) at 100°C for 10 min and loaded onto an SDS-polyacrylamide gel for western blot to detect binding efficiency.

Generation of Deletion/Truncation Constructs of DMAD, Wds, and Pc

Site-directed mutagenesis kit (NEB) was used to generate series of deletion constructs from full-length cDNA from DMAD, Wds and Pc. All constructs were subjected to Sanger sequencing to confirm the correct insertion. DMAD and Wds constructs were cloned into pAc 5.1 vector and expressed in *Drosophila* S2 cells for co-immunoprecipitation experiments. Pc deletion constructs were cloned into pFASTBAC plasmid for Baculovirus production (LakePharma).

Fluorescence-Based Pc-6mA Binding Assays

The 6-carboxy-fluorescein (FAM)-labeled control or 6mA-modified DNA probes were annealed with their reverse complementary strand to form double-stranded DNA oligos. Various concentrations (0.01-1 μM) of Pc full length or C-terminal truncation recombinant proteins were incubated with 2 nM DNA oligos for 15 min at room temperature in nicking buffer (10mM Tris-Cl pH8.0, 1mM EDTA, 0.1% BSA). Fluorescence polarization measurements were carried out at 25°C on a Synergy 4 microplate reader (BioTek). Curves were fit individually using GraphPad Prism 7.0 software (GraphPad Software). Binding constants (K_d) were calculated described previously (Hashimoto et al., 2014).

In Vitro 6mA Demethylation Assay

Double stranded control and 6mA-modified DNA oligos were mixed with DMAD C-terminal catalytic domain (aa 1657-2860) in reaction buffer containing 50mM HEPES (pH8.0), 2mM ascorbate, 1mM α-KG and 150 μM Fe at room temperature for 3 hr. The reactions were stopped by proteinase K digestion at 50°C for 2 hr. The purified DNA samples were subjected to 6mA dot blots and ImageJ quantification.

DpnI Digestion and qPCR

DpnI digestion and qPCR was conducted as previously described (Luo et al., 2016). Briefly, restriction enzyme digestion was performed by treating 1 ug of genomic DNA with 5 uL of 5 U/uL DpnI restriction enzyme (NEB) at 37°C for 1 hr. The digested DNA and non-digested DNA (5 ng) were subjected to qPCR using FastStart SYBR Green Master kit. The restriction enzyme digestion method takes advantage of the 6mA-sensitive restriction enzyme DpnI that preferentially cleaves methylated adenine at GATC/CATC/GATG sites. Equal amounts of DpnI-digested DNA and undigested control DNA were subjected to qPCR analyses with primers targeting 6mA dynamic regions identified by 6mA-IP. The percentage of 6mA in either control or DMAD-KD can be assessed by qPCR amplification and normalized to undigested DNA control (digested/undigested). Loci with lower 6mA modification in controls would hinder DpnI digestion, resulting in higher PCR fold changes than DMAD-KD samples.

Library Preparation and High-Throughput Sequencing

Enriched DNA from 6mA-IP and ChIP were subjected to library construction using the NEBNext ChIP-Seq Library Prep Reagent Set from Illumina according the manufacturer's protocol. Briefly, 25 ng of input genomic DNA or experimental enriched DNA were utilized

for each library construction. DNA fragments (150-300 bp) were selected by AMPure XP Beads (Beckman Coulter) after adaptor ligation. An Agilent 2100 BioAnalyzer was used to quantify amplified DNA and qPCR was applied to accurately quantify library concentration. 20 pM diluted libraries were used for sequencing. 50-cycle single-end sequencings were performed using Illumina HISeq 2000. Image processing and sequence extraction were done using the standard Illumina Pipeline. RNA-seq libraries were generated from duplicated samples per condition using the Illumina TruSeq RNA Sample Preparation Kit v2 following manufacturer's protocol. RNA-seq libraries were sequenced as 50-cycle pair-end runs using Illumina HISeq 2000.

Bioinformatics Analyses

Bioinformatics analyses for ChIP-seq and 6mA-IP-seq were conducted as described previously (Szulwach et al., 2011; Yao et al., 2014). Briefly, FASTQ sequence files were aligned to the dm3 reference genome using Bowtie v1.1.2 (Langmead et al., 2009). Peaks were identified by Model-based Analysis of ChIP-Seq (MACS) software (Zhang et al., 2008). Ngsplot software was used to calculate and plot unique 6mA and ChIP-seq mapped reads various genomic regions and generated heatmaps (Shen et al., 2014). Annotation and motif analysis were performed using the HOMER (Heinz et al., 2010) suite. RNA-seq reads were aligned using Tophat v2.0.8 (Trapnell et al., 2012) and differential RPKM expression values were extracted using cuffdiff v2.2.1 (Trapnell et al., 2012). Genomic interval overlapping analyses were performed using Bedtools (Quinlan and Hall, 2010). GO analyses were performed by The Database for Annotation, Visualization and Integrated Discovery (DAVID) v6.7 (Huang da et al., 2009) and GO Consortium (Ashburner et al., 2000). Gain-of-6mA regions in BG3C2 cells were identified by published computational algorithm (R/Bioconductor package DSS) that implements a series of differential methylation detection algorithms based on the dispersion shrinkage method followed by Wald statistical test to rigorously interrogate the 6mA differential regions between replicated samples of control and DMAD-KD group (Feng et al., 2014). Top-ranked gain-of-6mA regions were further intersected with 6mA peaks identified by MACS with FDR < 0.05. modENCODE ChIP-chip regions were generated by CisGenome (Ji et al., 2006). Transposon expression was analyzed by TEtranscripts (Jin et al., 2015). Gain-of-6mA regions in brains were identified by bedtools (Quinlan and Hall, 2010). MNase-seq data were analyzed using dynamic analysis of nucleosome position and occupancy by sequencing (DANPOS) (Chen et al., 2013).

Sample Size and Statistics

Fly brain RNA-seq experiments were performed using two control and DMAD null pooled samples as biological replicates. Each pooled sample contained 500-1000 fly brains per sample per genotype. Fly brains were dissected from heads to avoid 6mA contamination from bacteria. BG3C2 RNA-seq samples were performed in triplicate from control and DMAD-KD cells. Brain RNA-seq were performed in duplicates. Differential expression analyses were performed by cuffdiff (Trapnell et al., 2012). 6mA-IP in control and DMAD null brains were performed using biological replicates from 1000 pooled fly brains per sample per genotype. 6mA-IP experiments in BG3C2 cells were performed in duplicate using control and DMAD-KD cells. Control and DMAD-KD replicate 1 was sequenced twice serving as a technical replicate. DMAD ChIP-seq experiments in BG3C2 cells were performed in triplicate. All replicated samples from each condition were merged for downstream bioinformatic analyses. Pearson's Chi-square tests with Yates' continuity correction and Welch Two Sample t tests were performed in R computational environment (http://www.r-project.org/). Student's t tests were performed in Graphpad Prism (http://www.graphpad.com/scientific-software/prism/).

DATA AND SOFTWARE AVAILABILITY

The accession number for the ChIP-seq, RNA-seq, and 6mA-IP-seq reported in this paper is GEO: GSE67855.

Original imaging data have been deposited to Mendeley Data and are available at https://data.mendeley.com/datasets/wy3jshg29m/3.

3. W. Kim, S. Kook, D. J. Kim, C. Teodorof, W. K. Song, *J. Biol. Chem.* **279**, 8333 (2004).
4. V. Giambra *et al.*, *Mol. Cell. Biol.* **28**, 6123 (2008).
5. F. E. Garrett *et al.*, *Mol. Cell. Biol.* **25**, 1511 (2005).
6. W. A. Dunnick *et al.*, *J. Exp. Med.* **206**, 2613 (2009).
7. M. Cogné *et al.*, *Cell* **77**, 737 (1994).
8. J. P. Manis *et al.*, *J. Exp. Med.* **188**, 1421 (1998).
9. A. G. Bébin *et al.*, *J. Immunol.* **184**, 3710 (2010).
10. E. Pinaud *et al.*, *Immunity* **15**, 187 (2001).
11. C. Vincent-Fabert *et al.*, *Blood* **116**, 1895 (2010).
12. R. Wuerffel *et al.*, *Immunity* **27**, 711 (2007).
13. Z. Ju *et al.*, *J. Biol. Chem.* **282**, 35169 (2007).
14. H. Duan, H. Xiang, L. Ma, L. M. Boxer, *Oncogene* **27**, 6720 (2008).
15. M. Gostissa *et al.*, *Nature* **462**, 803 (2009).
16. C. Chauveau, M. Cogné, *Nat. Genet.* **14**, 15 (1996).
17. C. Chauveau, E. Pinaud, M. Cogne, *Eur. J. Immunol.* **28**, 3048 (1998).
18. M. A. Sepulveda, F. E. Garrett, A. Price-Whelan, B. K. Birshtein, *Mol. Immunol.* **42**, 605 (2005).
19. E. Pinaud, C. Aupetit, C. Chauveau, M. Cogné, *Eur. J. Immunol.* **27**, 2981 (1997).
20. A. A. Khamlichi *et al.*, *Blood* **103**, 3828 (2004).
21. R. Shinkura *et al.*, *Nat. Immunol.* **4**, 435 (2003).
22. A. Yamane *et al.*, *Nat. Immunol.* **12**, 62 (2011).
23. M. Liu *et al.*, *Nature* **451**, 841 (2008).
24. J. Stavnezer, J. E. Guikema, C. E. Schrader, *Annu. Rev. Immunol.* **26**, 261 (2008).
25. S. Duchez *et al.*, *Proc. Natl. Acad. Sci. U.S.A.* **107**, 3064 (2010).
26. T. K. Kim *et al.*, *Nature* **465**, 182 (2010).

Acknowledgments: We thank T. Honjo for providing AID$^{-/-}$ mice and F. Lechouane for sorted B cells DNA samples. We are indebted to the cell sorting facility of Limoges University for excellent technical assistance in cell sorting. This work was supported by grants from Association pour la Recherche sur le Cancer, Ligue Nationale contre le Cancer, Cancéropôle Grand Sud-Ouest, Institut National du Cancer, and Région Limousin. The data presented in this paper are tabulated here and in the supplementary materials.

Supplementary Materials

www.sciencemag.org/cgi/content/full/science.1218692/DC1
Materials and Methods
Figs. S1 to S4
Tables S1 and S2
References (*27*–*30*)

4 January 2012; accepted 27 March 2012
Published online 26 April 2012;
10.1126/science.1218692

Quantitative Sequencing of 5-Methylcytosine and 5-Hydroxymethylcytosine at Single-Base Resolution

Michael J. Booth,[1]* Miguel R. Branco,[2,3]* Gabriella Ficz,[2] David Oxley,[4] Felix Krueger,[5] Wolf Reik,[2,3]† Shankar Balasubramanian[1,6,7]†

5-Methylcytosine can be converted to 5-hydroxymethylcytosine (5hmC) in mammalian DNA by the ten-eleven translocation (TET) enzymes. We introduce oxidative bisulfite sequencing (oxBS-Seq), the first method for quantitative mapping of 5hmC in genomic DNA at single-nucleotide resolution. Selective chemical oxidation of 5hmC to 5-formylcytosine (5fC) enables bisulfite conversion of 5fC to uracil. We demonstrate the utility of oxBS-Seq to map and quantify 5hmC at CpG islands (CGIs) in mouse embryonic stem (ES) cells and identify 800 5hmC-containing CGIs that have on average 3.3% hydroxymethylation. High levels of 5hmC were found in CGIs associated with transcriptional regulators and in long interspersed nuclear elements, suggesting that these regions might undergo epigenetic reprogramming in ES cells. Our results open new questions on 5hmC dynamics and sequence-specific targeting by TETs.

5-Methylcytosine (5mC) is an epigenetic DNA mark that plays important roles in gene silencing and genome stability and is found enriched at CpG dinucleotides (*1*). In metazoa, 5mC can be oxidized to 5-hydroxymethylcytosine (5hmC) by the ten-eleven translocation (TET) enzyme family (*2*, *3*). 5hmC may be an intermediate in active DNA demethylation but could also constitute an epigenetic mark per se (*4*). Levels of 5hmC in genomic DNA can be quantified with analytical methods (*2*, *5*, *6*) and mapped through the enrichment of 5hmC-containing DNA fragments that are then sequenced (*7*–*13*). Such approaches have relatively poor resolution and give only relative quantitative information. Single-nucleotide sequencing of 5mC has been performed by using bisulfite sequencing (BS-Seq), but this method cannot discriminate 5mC from 5hmC (*14*, *15*). Single-molecule real-time sequencing (SMRT) can detect derivatized 5hmC in genomic DNA (*16*). However, enrichment of 5hmC-containing DNA fragments is required, which causes loss of quantitative information (*16*). Furthermore, SMRT has a relatively high rate of sequencing errors (*17*), and the peak calling of modifications is imprecise (*16*). Protein and solid-state nanopores can resolve 5mC from 5hmC and have the potential to sequence unamplified DNA (*18*, *19*).

We observed the decarbonylation and deamination of 5-formylcytosine (5fC) to uracil (U) under bisulfite conditions that would leave 5mC unchanged (Fig. 1A and supplementary text). Thus, 5hmC sequencing would be possible if 5hmC could be selectively oxidized to 5fC and then converted to U in a two-step procedure (Fig. 1B). Whereas BS-Seq leads to both 5mC and 5hmC being detected as Cs, this "oxidative bisulfite" sequencing (oxBS-Seq) approach would yield Cs only at 5mC sites and therefore allow us to determine the amount of 5hmC at a particular nucleotide position by subtraction of this readout from a BS-Seq one (Fig. 1C).

Specific oxidation of 5hmC to 5fC (table S1) was achieved with potassium perruthenate ($KRuO_4$). In our reactivity studies on a synthetic 15-nucleotide oligomer single-stranded DNA (ssDNA) containing 5hmC, we established conditions under which $KRuO_4$ reacted specifically with the primary alcohol of 5hmC (Fig. 2A). Fifteen-nucleotide oligomer ssDNA that contained C or 5mC did not show any base-specific reactions with $KRuO_4$ (fig. S1, A and B). For 5hmC in DNA, we only observed the aldehyde (5fC) and not the carboxylic acid (*20*), even with a moderate excess of oxidant. The $KRuO_4$ oxidation can oxidize 5hmC in samples presented as double-stranded DNA (dsDNA), with an initial denaturing step before addition of the oxidant; this results in a quantitative conversion of 5hmC to 5fC (Fig. 2B).

To test the efficiency and selectivity of the oxidative bisulfite method, three synthetic dsDNAs containing either C, 5mC, or 5hmC were each oxidized with $KRuO_4$ and then subjected to a conventional bisulfite conversion protocol. Sanger sequencing revealed that 5mC residues did not convert to U, whereas both C and 5hmC residues did convert to U (fig. S2). Because Sanger sequencing is not quantitative, to gain a more accurate measure of the efficiency of transforming 5hmC to U, Illumina (San Diego, California) sequencing was carried out on the synthetic DNA containing 5hmC (122-nucleotide oligomer) after oxidative bisulfite treatment. An overall 5hmC-to-U conversion level of 94.5% was observed (Fig. 2C and fig. S14). The oxidative bisulfite protocol was also applied to a synthetic dsDNA that contained multiple 5hmC residues (135-nucleotide oligomer) in a range of different contexts that showed a similarly high conversion efficiency (94.7%) of 5hmC to U (Fig. 2C and fig. S14). Last, the $KRuO_4$ oxidation was carried out on genomic DNA and showed through mass spectrometry a quantitative conversion of 5hmC to

[1]Department of Chemistry, University of Cambridge, Cambridge CB2 1EW, UK. [2]Epigenetics Programme, Babraham Institute, Cambridge CB22 3AT, UK. [3]Centre for Trophoblast Research, University of Cambridge, Cambridge CB2 3EG, UK. [4]Proteomics Research Group, Babraham Institute, Cambridge CB22 3AT, UK. [5]Bioinformatics Group, Babraham Institute, Cambridge CB22 3AT, UK. [6]School of Clinical Medicine, University of Cambridge, Cambridge CB2 0SP, UK. [7]Cancer Research UK, Cambridge Research Institute, Li Ka Shing Centre, Cambridge CB2 0RE, UK.

*These authors contributed equally to this work.
†To whom correspondence should be addressed. E-mail: wolf.reik@babraham.ac.uk (W.R.); sb10031@cam.ac.uk (S.B.)

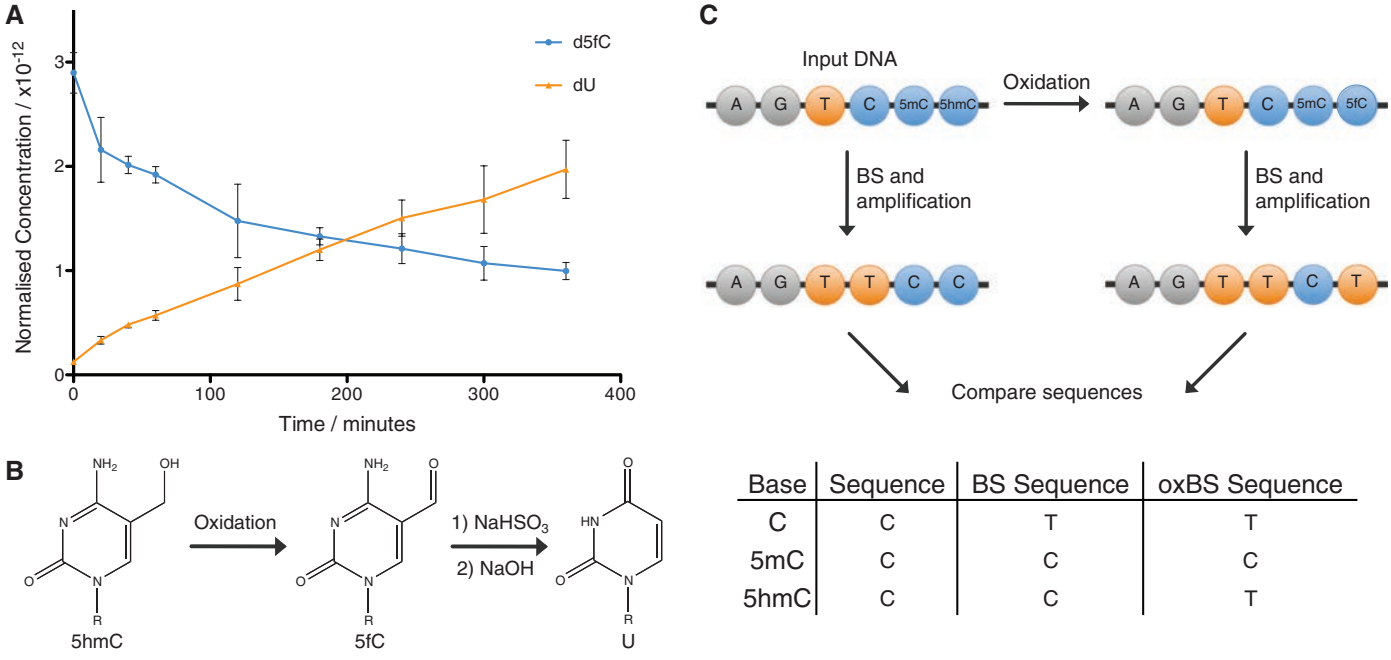

Fig. 1. A method for single-base resolution sequencing of 5hmC. (**A**) Reaction of 2′-deoxy-5-formylcytidine (d5fC) with $NaHSO_3$ (bisulfite) quenched by NaOH at different time points and then analyzed with high-performance liquid chromatography (HPLC). Data are mean ± SD of three replicates. (**B**) Oxidative bisulfite reaction scheme: oxidation of 5hmC to 5fC followed by bisulfite treatment and NaOH to convert 5fC to U. The R group is DNA. (**C**) Diagram and table outlining the BS-Seq and oxBS-Seq techniques.

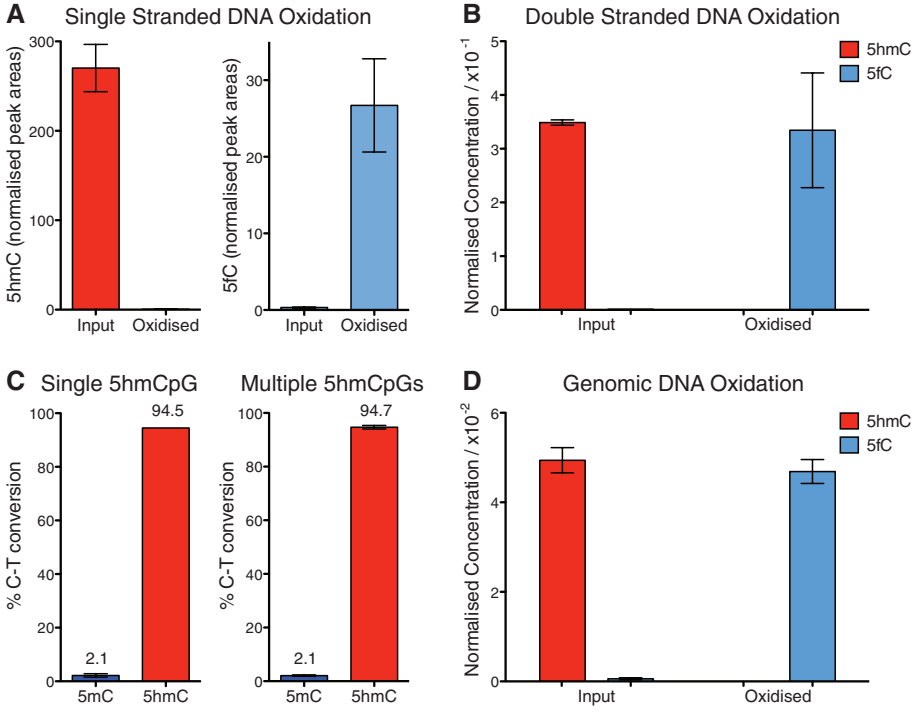

Fig. 2. Quantification of 5hmC oxidation. (**A**) Levels of 5hmC and 5fC (normalized to T) in a 15-nucleotide oligomer ssDNA oligonucleotide before and after $KRuO_4$ oxidation, measured with mass spectrometry. (**B**) Levels of 5hmC and 5fC (normalized to 5mC) in a 135-nucleotide oligomer dsDNA fragment before and after $KRuO_4$ oxidation. (**C**) C-to-T conversion levels as determined by means of Illumina sequencing of two dsDNA fragments containing either a single 5hmCpG (122-nucleotide oligomer) or multiple 5hmCpGs (135-nucleotide oligomer) after oxidative bisulfite treatment. 5mC was also present in these strands. (**D**) Levels of 5hmC and 5fC (normalized to 5mC in primer sequence) in ES cell DNA measured before and after oxidation. Data are mean ± SD.

5fC (Fig. 2D), with no detectable degradation of C (fig. S1C). Thus, the oxidative bisulfite protocol specifically converts 5hmC to U in DNA, leaving C and 5mC unchanged, enabling quantitative, single-nucleotide-resolution sequencing on widely available platforms.

We then used oxBS-Seq to quantitatively map 5hmC at high resolution in the genomic DNA of mouse embryonic stem (ES) cells. We chose to combine oxidative bisulfite with reduced representation bisulfite sequencing (RRBS) (*21*), which allows deep, selective sequencing of a fraction of the genome that is highly enriched for CpG islands (CGIs). We generated RRBS and oxidative RRBS (oxRRBS) data sets, achieving an average sequencing depth of ~120 reads per CpG, which when pooled yielded an average of ~3300 methylation calls per CGI (fig. S3). After applying depth and breadth cutoffs (supplementary materials, materials and methods), 55% (12,660) of all CGIs (*22*) were covered in our data sets.

To identify 5hmC-containing CGIs, we tested for differences between the RRBS and oxRRBS data sets using stringent criteria, yielding a false discovery rate of 3.7% (supplementary materials, materials and methods). We identified 800 5hmC-containing CGIs, which had an average of 3.3% (range of 0.2 to 18.5%) CpG hydroxymethylation (Fig. 3, A and B). We also identified 4577 5mC-containing CGIs averaging 8.1% CpG methylation (Fig. 3B). We carried out sequencing on an independent biological duplicate sample of the same ES cell line but at a different passage

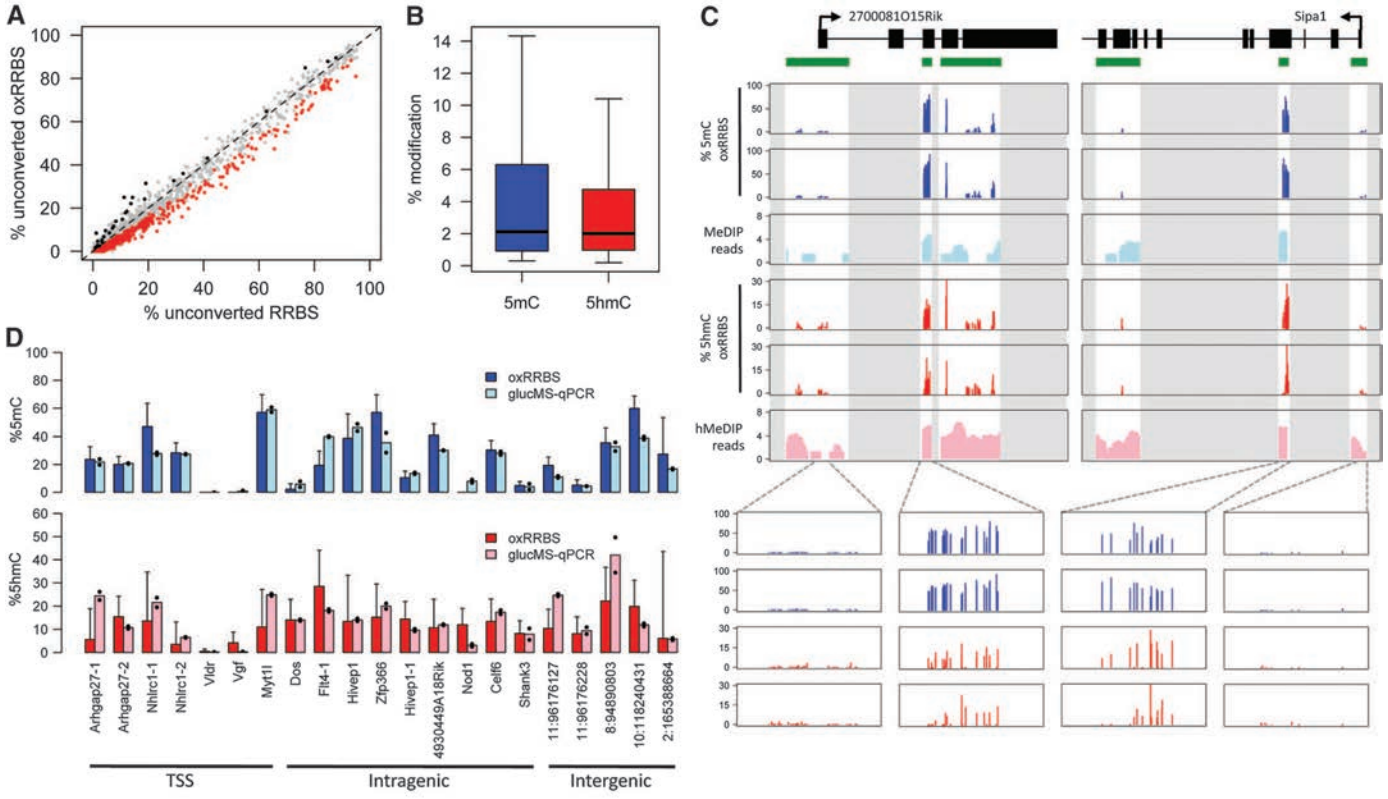

Fig. 3. Quantification of 5mC and 5hmC levels at CGIs by means of oxRRBS. (**A**) Fraction of unconverted cytosines per CGI; 5hmC-containing CGIs (red) have a statistically significant lower fraction in the oxRRBS data set; a false discovery rate of 3.7% was estimated from the CGIs with the opposite pattern (black). (**B**) 5mC and 5hmC levels within CGIs with significant levels of the respective modification. (**C**) Examples of genomic RRBS and oxRRBS profiles overlapped with (h)MeDIP-Seq profiles (*7*). Green bars represent CGIs; data outside CGIs were masked (gray areas). Each bar in the oxRRBS tracks represents a single CpG (in either DNA strand). (**D**) 5mC and 5hmC levels at selected MspI sites were validated through glucMS-qPCR. OxRRBS data are percentage ± 95% confidence interval. Mean glucMS-qPCR values are shown, with the black dots representing individual replicates.

number, which according to mass spectrometry had reduced levels of 5hmC (0.10 versus 0.16% of all Cs), and consistently we found fewer 5hmC-containing CGIs (supplementary text). 5hmC-containing CGIs present in both samples showed good quantitative reproducibility (fig. S5). In non-CpG contexts, we found very few CGIs (71) with levels of 5mC above the bisulfite conversion error (0.2%) (fig. S9) and no CGIs with detectable levels of 5hmC.

Genes associated with 5mC-containing CGIs included *Dazl*, which is known to be methylated in ES cells (fig. S7) (*23*). Similarly, we found that *Zfp64* and *Ecat1* had significant levels of 5hmC (*7*). Genes with >5% 5hmC at transcription start site (TSS) CGIs were associated with gene ontology terms related to transcription factor activity—and in particular were enriched in developmentally relevant genes encoding for Homeobox-containing proteins (such as *Irx4*, *Gbx1*, and *Hoxc4*). To validate our method, we quantified 5hmC and 5mC levels at 21 CGIs containing MspI restriction sites by means of glucosylation-coupled methylation-sensitive quantitative polymerase chain reaction (glucMS-qPCR) (Fig. 3D) (*24*). We found a good correlation between the quantification with oxRRBS and glucMS-qPCR [correlation coefficient (r) = 0.86,

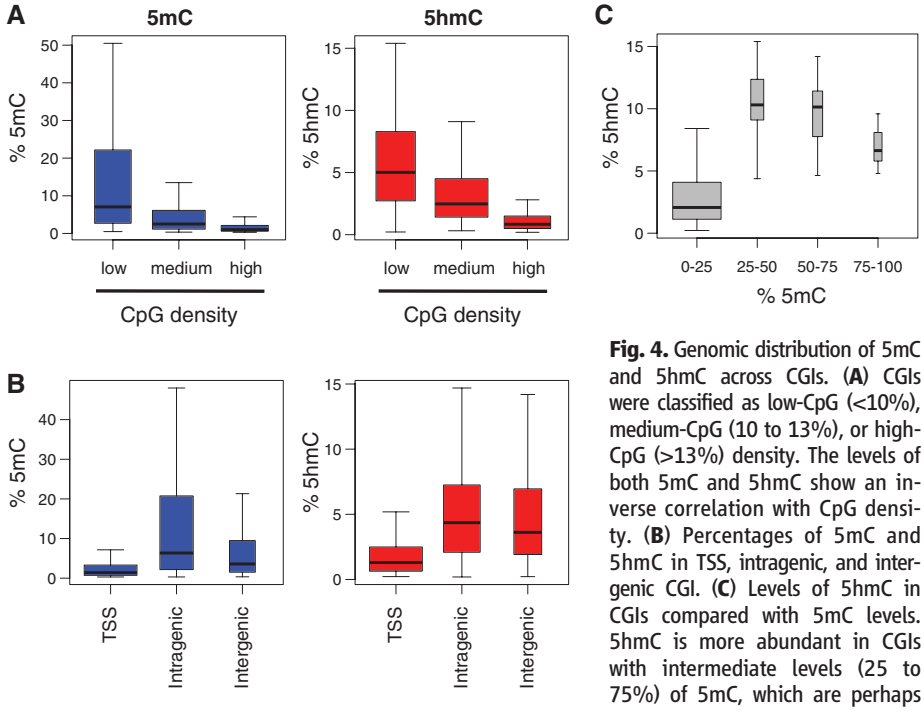

Fig. 4. Genomic distribution of 5mC and 5hmC across CGIs. (**A**) CGIs were classified as low-CpG (<10%), medium-CpG (10 to 13%), or high-CpG (>13%) density. The levels of both 5mC and 5hmC show an inverse correlation with CpG density. (**B**) Percentages of 5mC and 5hmC in TSS, intragenic, and intergenic CGI. (**C**) Levels of 5hmC in CGIs compared with 5mC levels. 5hmC is more abundant in CGIs with intermediate levels (25 to 75%) of 5mC, which are perhaps more epigenetically plastic. For all boxplots, the width of the box is proportional to the amount of data within that group.

$P = 5 \times 10^{-7}$ and $r = 0.52$, $P = 0.01$ for 5mC and 5hmC, respectively], showing that oxRRBS reliably measures 5hmC at individual CpGs. We also found a good correlation between oxRRBS and our previously published (hydroxy)methylated DNA immunoprecipitation sequencing [(h)MeDIP-Seq] data sets (fig. S8) (*7*).

Across CGIs, both 5mC and 5hmC levels are inversely correlated with CpG density, and intragenic and intergenic CGIs contain higher levels of either modification than those overlapping TSSs (Fig. 4, A and B, and fig. S6) (*13*, *22*). TET1 is enriched at TSSs, and thus, a high turnover of 5mC and 5hmC that would keep the steady-state levels low at these sites has been suggested (*9*). Non-TSS CGIs, however, appear to accumulate substantial amounts of both marks, suggesting reduced turnover in these regions. We find that the highest levels of 5hmC are found at CGIs with intermediate levels (25 to 75%) of 5mC (Fig. 4C and fig. S6). Although low-5mC CGIs have reduced potential for 5hmC generation and/or are subjected to a high turnover, high-5mC CGIs are perhaps protected from extensive TET-mediated oxidation, thus stabilizing methylation. Intermediate-5mC CGIs are therefore potentially more epigenetically plastic, given the relatively high abundance of both marks.

Most TSS CGIs (98%) have less than 10% 5mC, as well as low 5hmC, and these are associated with higher transcription levels than average (fig. S10). Within this narrow window, we find a mild negative correlation between transcription and both 5mC and 5hmC levels (fig. S10). At higher 5mC levels, there are insufficient CGIs to obtain a statistically significant result, and it remains possible that here the epigenetic balance between 5mC and 5hmC plays an important transcriptional role, as we previously suggested (*7*).

Last, we quantified 5mC and 5hmC levels at two classes of retrotransposons [long interspersed nuclear element–1 (LINE1) and intracisternal A-particle (IAP)] using two approaches: aligning the oxRRBS reads to the respective consensus sequences and combining oxidative bisulfite with MassARRAY technology (Sequenom, San Diego, California) (fig. S11). We find that LINE1 elements display a considerable amount of 5hmC (approximately 5%), as previously suggested through (h)MeDIP-Seq (*7*). IAPs, on the other hand, have low or no 5hmC. Because LINE1 elements are reprogrammed during preimplantation development whereas IAPs are resistant to this process (*25*), this suggests a possible involvement of 5hmC in the demethylation of specific repeat classes.

The oxBS-Seq method reliably maps and quantifies both 5mC and 5hmC at the single-nucleotide level. Owing to the fundamental mechanism of oxBS-Seq, the approach is compatible with any sequencing platform. In ES cells, we found that in CGIs 5hmC is exclusive to CpG dinucleotides and that it accumulates at intragenic, low-CpG-density CGIs, which tend to have intermediate levels of 5mC and may be particularly epigenetically plastic.

References and Notes

1. A. M. Deaton, A. Bird, *Genes Dev.* **25**, 1010 (2011).
2. M. Tahiliani *et al.*, *Science* **324**, 930 (2009).
3. S. Ito *et al.*, *Nature* **466**, 1129 (2010).
4. M. R. Branco, G. Ficz, W. Reik, *Nat. Rev. Genet.* **13**, 7 (2012).
5. S. Kriaucionis, N. Heintz, *Science* **324**, 929 (2009).
6. M. Münzel *et al.*, *Angew. Chem. Int. Ed.* **49**, 5375 (2010).
7. G. Ficz *et al.*, *Nature* **473**, 398 (2011).
8. W. A. Pastor *et al.*, *Nature* **473**, 394 (2011).
9. H. Wu *et al.*, *Genes Dev.* **25**, 679 (2011).
10. S. G. Jin, X. Wu, A. X. Li, G. P. Pfeifer, *Nucleic Acids Res.* **39**, 5015 (2011).
11. C. X. Song *et al.*, *Nat. Biotechnol.* **29**, 68 (2011).
12. K. Williams *et al.*, *Nature* **473**, 343 (2011).
13. Y. Xu *et al.*, *Mol. Cell* **42**, 451 (2011).
14. Y. Huang *et al.*, *PLoS ONE* **5**, e8888 (2010).
15. C. Nestor, A. Ruzov, R. Meehan, D. Dunican, *Biotechniques* **48**, 317 (2010).
16. C. X. Song *et al.*, *Nat. Methods* **9**, 75 (2012).
17. J. Eid *et al.*, *Science* **323**, 133 (2009).
18. E. V. Wallace *et al.*, *Chem. Commun. (Camb.)* **46**, 8195 (2010).
19. M. Wanunu *et al.*, *J. Am. Chem. Soc.* **133**, 486 (2010).
20. G. Green, W. P. Griffith, D. M. Hollinshead, S. V. Ley, M. Schroder, *J. Chem. Soc. Perkin Trans. 1* **1**, 681 (1984).
21. A. Meissner *et al.*, *Nature* **454**, 766 (2008).
22. R. S. Illingworth *et al.*, *PLoS Genet.* **6**, e1001134 (2010).
23. J. Borgel *et al.*, *Nat. Genet.* **42**, 1093 (2010).
24. S. M. Kinney *et al.*, *J. Biol. Chem.* **286**, 24685 (2011).
25. N. Lane *et al.*, *Genesis* **35**, 88 (2003).

Acknowledgments: We thank T. Green and R. Rodriguez for helpful discussions and J. Webster for help with mass spectrometry. We thank the Biotechnology and Biological Sciences Research Council (BBSRC) for a studentship (M.J.B.). The W.R. lab is supported by BBSRC, Medical Research Council, the Wellcome Trust, European Union EpiGeneSys, and BLUEPRINT. The S.B. lab is supported by core funding from Cancer Research UK. M.J.B. and S.B. are inventors on provisional applications filed for U.S. patents on oxBS-Seq (patent applications US61/605702; US61/641134; US61/623461; and US61/513356). OxRRBS data are deposited in the European Molecular Biology Laboratory–European Bioinformatics Institute ArrayExpress Archive (http://www.ebi.ac.uk/arrayexpress) under the accession number E-MTAB-1042. S.B. is an advisor to Illumina.

Supplementary Materials

www.sciencemag.org/cgi/content/full/science.1220671/DC1
Materials and Methods
Supplementary Text
Figs. S1 to S15
Tables S1 and S2
References (*26*–*40*)

16 February 2012; accepted 13 April 2012
Published online 26 April 2012;
10.1126/science.1220671

Science
AAAS

Quantitative Sequencing of 5-Methylcytosine and 5-Hydroxymethylcytosine at Single-Base Resolution

Michael J. Booth, Miguel R. Branco, Gabriella Ficz, David Oxley, Felix Krueger, Wolf Reik and Shankar Balasubramanian (April 26, 2012)

Science **336** (6083), 934-937. [doi: 10.1126/science.1220671] originally published online April 26, 2012

Editor's Summary

Distinguishing Epigenetic Marks

Methylation of the cytosine base in eukaryotic DNA (5mC) is an important epigenetic mark involved in gene silencing and genome stability. Methylated cytosine can be enzymatically oxidized to 5-hydroxymethylcytosine (5hmC), which may function as a distinct epigenetic mark––possibly involved in pluripotency––and it may also be an intermediate in active DNA demethylation. To be able to detect 5hmC genome-wide and at single-base resolution, **Booth *et al.*** (p. 934, published online 26 April) developed a 5hmC sequencing chemistry that selectively oxidizes 5hmC to 5-formylcytosine and then to uracil while leaving 5mC unchanged. Using this method, mouse embryonic stem cell genomic D NA was sequenced to reveal that 5hmC is found enriched at intragenic CpG islands and long interspersed nuclear element–1 retrotransposons.

Science (print ISSN 0036-8075; online ISSN 1095-9203) is published weekly, except the last week in December, by the American Association for the Advancement of Science, 1200 New York Avenue NW, Washington, DC 20005.

REVIEW

Single-cell epigenomics: Recording the past and predicting the future

Gavin Kelsey,[1,2]*† Oliver Stegle,[3,4]*† Wolf Reik[1,2,5]†

Single-cell multi-omics has recently emerged as a powerful technology by which different layers of genomic output—and hence cell identity and function—can be recorded simultaneously. Integrating various components of the epigenome into multi-omics measurements allows for studying cellular heterogeneity at different time scales and for discovering new layers of molecular connectivity between the genome and its functional output. Measurements that are increasingly available range from those that identify transcription factor occupancy and initiation of transcription to long-lasting and heritable epigenetic marks such as DNA methylation. Together with techniques in which cell lineage is recorded, this multilayered information will provide insights into a cell's past history and its future potential. This will allow new levels of understanding of cell fate decisions, identity, and function in normal development, physiology, and disease.

The discovery and description of individual cells in the body has fascinated biologists and pathologists since the cell was discovered (*1*). With the advent of molecular cell biology, methods have been developed for measuring properties and functions of single cells at increasing resolution. This includes, among others, fluorescent protein reporters and single-molecule detection of RNA or DNA. Only recently however, have high-throughput sequencing methods allowed us more comprehensive access to genomic information in single cells. Hence, single-cell RNA sequencing has revealed how heterogeneous the transcriptome of individual cells can be within a seemingly homogeneous cell population or tissue, providing insights into cell identity, fate, and function in the context of both normal biology and pathology [Stubbington *et al.* (*2*) and Lein *et al.* (*3*)]. A few years from now, we likely will have access to total RNA, small and long noncoding RNA, and transcriptional initiation output of the transcriptome (in addition to the stable cytoplasmic component). The development of single-cell RNA sequencing was followed by single-cell genome sequencing, which has provided new insights into genomic stability and genomic variations that occur in physiology and in disease—for example, in cancer, reproductive medicine, or microbial genetics (*4*).

Epigenetics connects the genome with its functional output (Fig. 1). Various epigenetic marks have been described, ranging from DNA (such as DNA methylation) to histone modifications, which can affect the way the cell reads its genome and hence its transcriptional output. Transcription factors that bind to DNA can create or alter epigenetic states (e.g., open or closed chromatin and higher-order chromatin conformation), or their binding can be sensitive to preexisting epigenetic states. Some epigenetic marks can also be heritable from one cell generation to the next (during mitosis) or from one organism generation to the next [intergenerational or transgenerational epigenetic inheritance (*5*)]. However, there are key questions in epigenetics that can only be addressed by determining the epigenome in single cells. For example, how is transcriptional heterogeneity between cells connected with epigenetic heterogeneity (if it is), do changes in transcription precede or follow epigenetic marks when cells change their fate or function, and are epigenetic states better or worse identifiers of rare cell populations and transitional states than the transcriptome? The recent development of single-cell epigenomics methods is beginning to allow us to address these fundamental questions.

> ***"[T]oday we can probe the majority of epigenetic dimensions with single-cell resolution."***

Single-cell epigenome methods can identify open or closed chromatin, including nucleosome positioning (*6*–*11*). From these, one can infer the likelihood of certain transcription factors to bind or not bind to specific DNA sequences within individual cells, and methods are being developed that allow for assaying transcription factor binding directly—for example, single-cell chromatin immunoprecipitation sequencing (ChIP-seq). Thus, one can currently measure (albeit imperfectly) the heterogeneity in a cell population of key histone marks associated with transcriptional states, such as H3K4me3, which indicates active transcription, or H3K27me3, which is found on genes with a repressed transcriptional state (*12*). Functional states (such as transcriptional output) of the genome are also guided by the way the DNA in each cell is organized into higher-order chromatin, which can be determined by single-cell high-throughput chromosome conformation capture (Hi-C) (*13*). Finally, various DNA modifications—such as methylation (5mC), hydroxymethylation (5hmC), and formylcytosine (5fC)—can be located at the single-cell level by sequencing in most areas of the genome, including at single-nucleotide resolution (*14*–*18*). These modifications are part of the biological turnover of DNA methylation and are associated, for example, with transcriptional repression (5mC) or enhancers, including active ones (5hmC and 5fC). Hence, today we can probe the majority of epigenetic dimensions with single-cell resolution.

The techniques described above have been combined into single-cell multi-omics (*19*), which can reveal new connections between regulatory principles that operate in the individual layers (Figs. 1 and 2). Hence, genome sequencing together with transcriptome sequencing can reveal how genetic variation is related to transcriptional variation (*20*, *21*). Furthermore, genome-scale methylome sequencing coupled with the transcriptome (*22*, *23*) has identified widespread associations between epigenetic marks and transcriptional heterogeneity. The latest incarnation, triple-omics, combines genome, methylome, and transcriptome (*24*) assays and can reveal methylome, chromatin accessibility, and the transcriptome (*11*). Together with the development of multidimensional computational methods (*22*, *25*), these techniques are beginning to tease out intricate and unique cell- and locus-specific relationships between, say, methylation and nucleosome accessibility of a gene promoter and the transcriptional output of the gene (*11*).

Single-cell profiling of DNA modifications

Because epigenetic information comes in multiple forms—covalent modifications on DNA, posttranslational modifications of histones, chromatin accessibility and compaction, and higher-order conformation of chromosome domains—each layer of information requires a different biochemical approach to profile it. This has implications for the nature and quality of the information generated from single cells and for the ability to combine multiple measures from the same single cell in multi-omic applications. Depending on the type of question, it will be necessary to determine whether depth or breadth (many, many cells) is required for any specific study (Fig. 2).

Technically, DNA methylation has been the easiest to assay, building on well-established bisulphite chemistry (*26*). However, bisulphite treatment degrades DNA, preventing full-genome coverage and requiring an adaptation of bisulphite sequencing (BS-seq) to the single-cell level (*14*–*16*). BS-seq, by which unmodified cytosine is converted to thymine but 5mC remains unconverted (*26*), yields single-base precision in principle, with the advantage that both modified and unmodified sites are identified (*26*). Therefore, sites without

[1]Epigenetics Programme, Babraham Institute, Cambridge CB22 3AT, UK. [2]Centre for Trophoblast Research, University of Cambridge, Cambridge CB2 3EG, UK. [3]European Molecular Biology Laboratory, European Bioinformatics Institute, Wellcome Genome Campus, CB10 1SD Hinxton, Cambridge, UK. [4]European Molecular Biology Laboratory, Genome Biology Unit, Heidelberg 69117, Germany. [5]Wellcome Trust Sanger Institute, Cambridge CB10 1SA, UK.
*These authors contributed equally to this work. †Corresponding author. Email: gavin.kelsey@babraham.ac.uk (G.K.); oliver.stegle@ebi.ac.uk (O.S.); wolf.reik@babraham.ac.uk (W.R.)

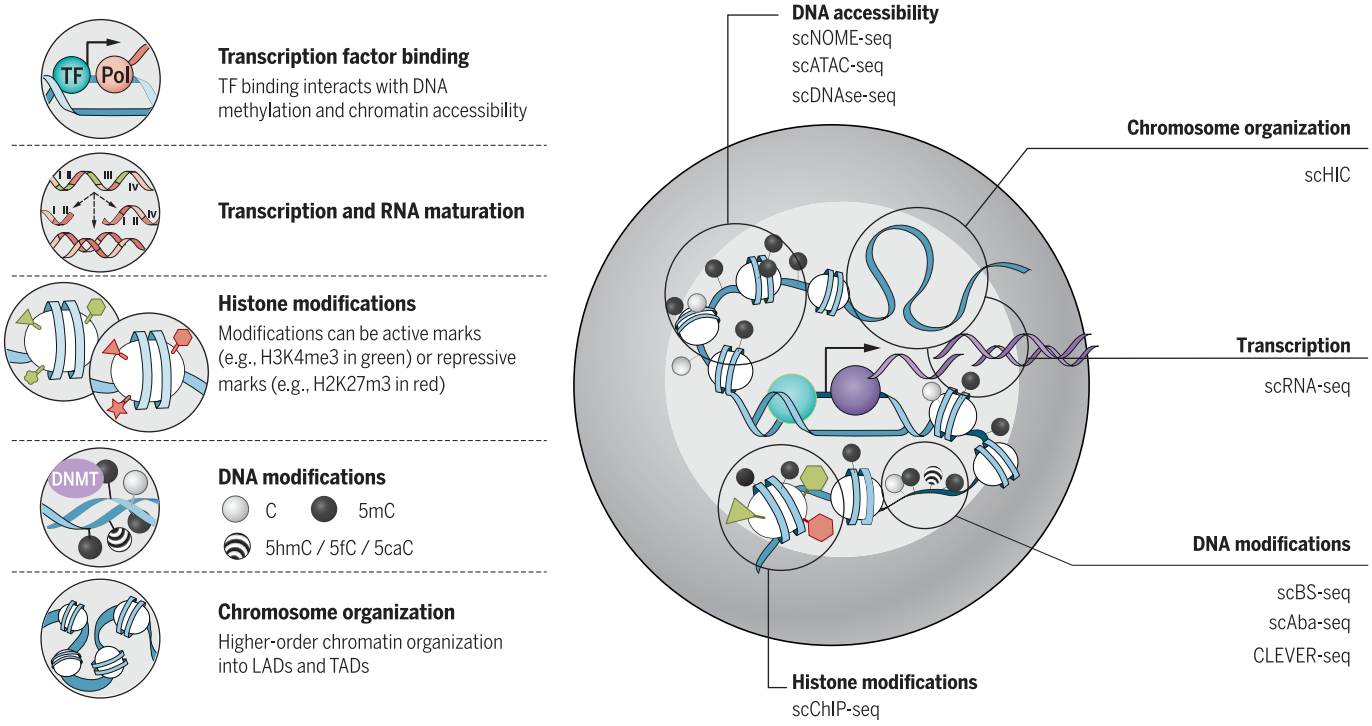

Fig. 1. Single-cell methods and heterogeneity of different molecular layers. (**Left**) Overview of different molecular layers that can be assayed using single-cell protocols. (**Right**) A cell with different layers of multi-omics measurements, as defined on the left. Concordance or heterogeneity respectively may exist between the different layers, and this can be recorded by single-cell sequencing and computationally evaluated.

information are not falsely assigned as unmethylated and, because of the general congruence of methylation over consecutive CpGs in many genomic contexts, missing sites can be imputed from relatively sparse data.

Current single-cell BS-seq (scBS-seq) protocols achieve a coverage up to ~40% (*15*), which means that for most loci the observed sequence reads will originate from only one chromosomal copy. Recent advances in performing single-cell methylation profiling with combinatorial indexing (*27, 28*) may mitigate some of these limitations while simultaneously offering scalability to thousands of cells in a single experiment (Fig. 2). Alternatively, because methylation state can determine whether particular restriction enzymes cleave their recognition sites, methods that use methylation-sensitive or dependent restriction enzymes could present an alternative to bisulphite-based methods (*29*).

Mapping the derivatives of 5mC in single cells has been particularly useful in preimplantation embryos, in which oxidation of 5mC contributes to the active demethylation of the paternal chromosomes (*30*). The pronounced strand bias in distribution between sister cells of these modifications along the same chromosome has provided high-resolution analysis of sister-chromatin exchange (*31*) and has been used as a lineage reconstruction tool (*17*), as well as mapping active demethylation in advance of expression at the promoters of developmentally important genes (*18*). Such advances have required alternative approaches, because 5mC cannot be discriminated from the less abundant 5hmC after bisulphite treatment, and the rarer derivatives 5fC and 5-carboxycytosine (5caC) are indistinguishable from unmodified cytosine.

Treatment with the CpG methylase M.SssI [methylase-assisted bisulphite sequencing, (MAB-seq)] (*31*) allows indirect detection of 5fC, together with 5caC, due to their retention as the only sites remaining susceptible to C to T conversion after bisulphite treatment. Careful control of the methylation reaction is needed to minimize false-positive calls, particularly for a rare modification such as 5fC, which is present at most at tens of thousands of CpG sites, compared with millions of CpGs modified by 5mC. 5hmC can be profiled in single cells by glucosylating 5hmC positions to generate recognition sites for the restriction endonuclease AbaS1 (scAba-seq) (*17*). This provides a positive readout of 5hmC, but, with the inclusion of multiple enzymatic reactions, there is an unknown false-negative rate, which might contribute to a range in the number of 5hmC positions recorded in single cells. 5fC can be detected in single cells by direct chemical labeling with the specific reactivity of malononitrile [chemical-labeling-enabled C-to-T conversion sequencing (CLEVER-seq)] (*18*). The adduct produced prevents normal pairing with G, such that labeled 5fC sites are read as T during polymerase chain reaction (PCR) amplification. In theory, this approach may allow for robust detection of modified bases on single-molecule sequencing platforms.

Combining methylation profiling into multi-omics approaches

scBS-seq can be combined with scRNA-seq through separation of nuclei from cell cytoplasm, separation of RNA and DNA for separate downstream reactions, or preamplification of RNA and DNA in the same cell lysate before splitting and parallel processing for genomic DNA amplification and cDNA library preparation (*22–24*). BS-seq coverage is sufficiently uniform to permit identification of chromosome aneuploidies or large CNVs from regional variations in read depth (*24*). Of note, similar to scRNA-seq protocols that use plate-based methods, scBS-seq can in principle be coupled with profiling of up to tens of cell-surface markers that can be assayed using fluorescence-activated cell sorting, an approach that has been applied in immunology [see Stubbington *et al.*, (*2*)].

Bisulphite sequencing also underlies the nucleosome occupancy and methylome (NOME) sequencing method, which enables information on nucleosome positioning and accessible chromatin to be inferred simultaneously with DNA methylation (*9–11*). Individual lysed cells are treated with M.CviPI, which methylates GpC sites in accessible DNA; then, following bisulphite treatment, methylated cytosines in a GpC context demarcate accessible DNA (linker regions and nucleosome-free DNA), while methylation is read from conversion events of CpGs. Because

both accessible and nonaccessible states are reported, missing information is not falsely assigned, which provides an advantage over other methods for chromatin accessibility. On the other hand, as a method that sequences the genome with no selectivity for open chromatin, high levels of sequencing may be needed to guarantee coverage of elements of interest.

Another potential limitation is the need to filter out C-C-G and G-C-G positions from the methylation data, which reduces the number of genome-wide cytosines that can be assayed compared with scBS-seq by ~50%. However, despite this filter, a large proportion of the loci in genomic regions with important regulatory roles, such as promoters and enhancers, can still be profiled using scNOME-seq–based methods (*11*). scNOME-seq has identified chromatin remodeling dynamics on the two parental alleles during preimplantation development, discriminating cis-regulatory elements open in all cells and promoters that diverge in accessibility between individual blastomeres, these being relatively enriched in gene ontology (GO) terms related to developmental processes and cell differentiation (*9*). Further enhancements of these data can be provided by incorporating transcriptome information from the same cell (Fig. 2) (*11*) to query the strength of coupling between DNA methylation, open chromatin, and transcriptional output.

Mapping functional chromatin states in single cells

A variety of assays have been adapted to profile chromatin states in single cells; these are predicated on enrichment-based strategies; thus, in principle, they have a lower sequencing overhead than scNOME-seq. Open chromatin can be identified by deoxyribonuclease I (DNase I) sensitivity, which was first adapted to the single-cell level in a low-throughput application able to detect an average of ~40,000 DNase I hypersensitive sites (DHSs) per cell (*6*). However, due to nonspecific signals throughout the genome, the false-discovery rate is high. Thus, previous knowledge of DHSs from bulk experiments is required to identify genuine DHSs, with the confidence of detection of proximal regulatory elements scaling with expression level of associated genes.

Higher-throughput applications have been developed for the assay for transposase-accessible chromatin sequencing (ATAC-seq), in which DNA accessibility is probed by the ability of the prokaryotic Tn5 transposase to insert sequencing adapters into accessible regions of the genome, in contrast to regions that are inaccessible, such as those interacting with a nucleosome. These approaches have used microfluidics to process single cells and introduce cell-identifying barcodes as part of the tagging process (*7*) or by combinatorial-cell barcoding (*8*) (Fig. 2), allowing parallel processing of a large number of samples (>10,000).

Throughput levels face a cost of reduced depth, as typically <10% of known promoters are represented in an individual scATAC-seq library. Sparseness of data limits analysis of cellular variation at individual regulatory elements. This may preclude ab initio identification of open chromatin sites, and the absence of open chromatin at a locus of interest in a single cell may reflect missing data. As well as reporting active regulatory elements governing hematopoietic differentiation, scATAC-seq has identified the evolution of regulatory elements during disease progression in acute myeloid leukemia (*32*). In addition, the ability of scATAC-seq to delineate the cis-regulatory landscapes of constituent cell types from a complex solid tissue has been demonstrated by isolating single nuclei from frozen samples of mouse forebrain (*33*).

"Technological advances for assaying epigenetic diversity at the single-cell level have gone hand-in-hand with computational methods for interpreting the data generated."

Posttranslational modifications of histones that correlate with chromatin activity states are conventionally mapped by ChIP-seq. Adapting ChIP-seq to extract this information from single cells presents additional problems of specificity and sensitivity, because it is dependent on antibody binding to pull down modified histones with associated DNA. Droplet approaches and cellular barcoding to label nuclei individually at the stage of micrococcal nuclease digestion (which fragments chromatin into nucleosomes) with immunoprecipitation on pools of cells and subsequent deconvolution of single-cell data after multiplex library sequencing allow thousands of single cells to be processed in single experiments (*12*) (Fig. 2). Yet, although ~50% of sequencing reads may fall within known peaks of H3K4me3 enrichment (the archetypal mark of active promoters), only ~5% of known peaks are detected per cell, with data too sparse for productive de novo peak calling.

We shall inevitably see technical improvements in each of these chromatin profiling methods, as well as incorporating them into multi-omic approaches. A challenge is to extract RNA from cell lysates in a way that preserves both chromatin state and RNA integrity, but with the sparsity of data from current scATAC-seq, scDNase-seq, or scChIP-seq methods, attainment of parallel data on gene expression and chromatin state at specific loci is challenging, and processing increasing numbers of cells may be necessary to obtain sufficient convergent information. Any of the above methods in theory could be combined with bisulphite sequencing to investigate DNA methylation state, which is not to underestimate the technical challenges that may need to be overcome in adding the chemical steps involved in bisulphite treatment.

Readouts of gross chromatin organization in single cells

Higher orders of chromosome organization in interphase nuclei are represented by a number of configurations: topologically associated domains (TADs) divide the genome into structurally separate segments contained in loops and constrained by boundary elements, and lamin-associated domains (LADs) occupy the nuclear periphery. LADs have been probed at the single-cell level by Dam-ID, in which the Dam adenosine methyltransferase is fused with lamin B1 (a constituent of the nuclear lamina) and expressed in cells so that sites of interaction are mapped from sequence tags after DpnI digestion (*34*). Because LADs are megabase-scale chromosome domains, with 1100 to 1400 domains present in a typical cell, only a low rate of false negatives is expected. The extent of heterogeneity between cells thus allows a good measure of the numbers of constitutive and facultative LADs, as well as cooperativity between LADs; such data are not accessible from population-based approaches. Dam-ID methodology could be applied to any other protein interacting with DNA, such as chromatin remodelers and transcription factors. One caveat is that the false-negative rate will increase as the domain of interaction diminishes, or for proteins with very transient interactions.

Hi-C data measures the proximity of DNA sequences in three-dimensional (3D) space on the basis of ligation events in fixed nuclei. A variety of optimizations have been introduced to increase resolution of the data (*35*), as well as throughput (*36*, *37*), since the first report of a single-cell Hi-C method (*13*). Using haploid cells, single-cell Hi-C has allowed modeling of the 3D organization of all chromosomes in individual cells (*38*) and revealed how bulk-cell data obscures the dynamic reorganization of chromosome compartments during the cell cycle (*36*). Despite recent advances, the resolution of scHi-C methods remains insufficient to interrogate contacts between specific promoters and their enhancers, which awaits progress in miniaturizing approaches to promoter-capture Hi-C or complementation with functional experiments, such as epigenome editing (*39*).

Scalability and limitation of current methods

There are common challenges and limitations that apply to several single-cell epigenome methods. An important bottleneck is the currently limited capture rate (e.g., up to ~40% for scBS-seq), which means that even if libraries are sequenced to saturation, missing values are unavoidable (Fig. 3). Other potential drawbacks are low mappability rates (~20 to 30%) and high levels of PCR duplicates (*15*), in particular for deeply sequenced libraries (*16*), which need to be considered when analyzing the resulting data.

So far, epigenome-based methods tend to offer lower throughput than scRNA-seq, which can already be scaled to tens or hundreds of thousands of cells. Recent advances to perform single-cell methylation profiling, ATAC-seq, and Hi-C using combinatorial indexing (*8*, *28*, *37*) have narrowed

this gap. However, in particular, multi-omics methods that require a physical separation step of the RNA and DNA remain limited to medium-throughput analyses of hundreds of cells (Fig. 2). Another current challenge is to estimate and control for technical sources of variation. In single-cell transcriptomics, the level of technical noise can be estimated with spike-in standards, but such normalization strategies are not established for epigenome sequencing. A general strategy that can be useful are negative and positive controls—e.g., diluted bulk material used to create "pseudo cells" or control wells that combine one cell each from different species (*16*), which can be processed alongside each batch of single cells.

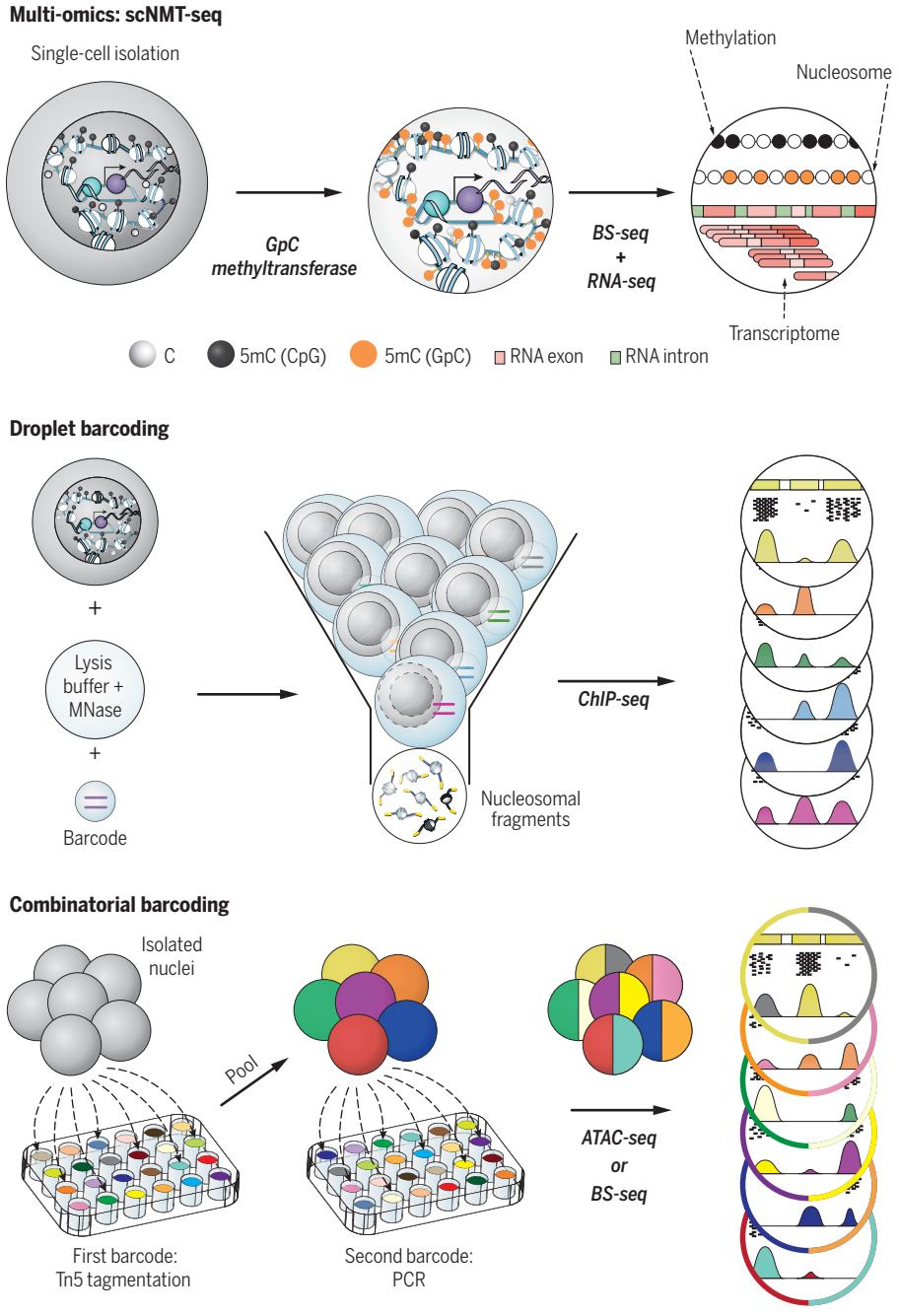

Fig. 2. Depth versus breadth: Multi-omics and cell-barcoding methods. Examples of different technical approaches are shown. (Top) Single-cell nucleosome, methylation, and transcription sequencing (scNMT-seq) (*11*) by which nucleosome accessibility, DNA methylation, and the transcriptome are read simultaneously at considerable depth in each cell; however, with individual cells processed in parallel but separately, cell numbers that can be currently analyzed in this way are limited to hundreds or thousands. (Middle) Barcoding chromatin in individual cells encapsulated in oil droplets, followed by pooling to bulk up material, enables thousands of cells to be processed while seeking to preserve signal-to-noise ratio (*12*). (Bottom) Combinatorial-cell barcoding (*8*, *64*), where readouts can be identified as coming from individual cells by unique combinations of barcodes present in each cell. This approach can be carried out on large numbers of cells (millions), but the depth of information per cell is limited.

Computational analysis to account for missing information using pooling strategies and imputation

Technological advances for assaying epigenetic diversity at the single-cell level have gone hand-in-hand with computational methods for interpreting the data generated (Fig. 3). A first critical step in the computational analysis is the appropriate normalization of the sequencing data while accounting for the typically high levels of noise observed. The sparse coverage of processed single-cell epigenome data sets requires careful consideration in downstream analyses.

Protocols vary in their coverage and whether missing data can be identified directly. For methods that use a bisulphite conversion step, the read coverage is independent of variation in DNA methylation, and hence missing data can be readily identified. For other methods, such as single-cell ATAC-seq, this can be more difficult because the absence or presence of sequence reads is the primary readout of the assay. Different strategies to address the low coverage in these data, such as aggregating read information within regions, by combining reads in consecutive sequence windows (*15*, *16*, *40*) or in annotated genomic contexts, such as promoter regions, enhancers and the like have been proposed. However, there are trade-offs between spatial resolution and coverage, parameters that may greatly affect downstream analyses.

Depending on the question, it may be advantageous to adjust for differences in global methylation, either at the whole-cell level or stratified by genomic context (*16*). A second strategy is to pool cells with similar epigenetic profiles, such as with an initial clustering step to then aggregate read information across cells within each cluster (*27*). These average profiles can offer high spatial resolution, however, at the cost that epigenetic diversity can only be studied at the level of the identified cell clusters (*24*). A third strategy comprises model-based approaches to impute missing information with predictive models. Such strategies have been proposed in the context of bulk epigenome profiles (*41*, *42*) and most recently have been generalised for imputing single-cell DNA methylation data (*25*). Additionally, we note that parallel data from multi-omics experiments will be associated with different patterns of missing data. Because of cost and experimental limitations, not all molecular layers will be assayed in each cell, and hence new computational methods need to handle heterogeneous designs to impute entire molecular layers.

Interrogating single-cell epigenome variation

Depending on the biological question at hand, several downstream analyses can be considered. Caution is required to consider the biological

sources of variation that one may expect in a given study. For example, the cell cycle is a dominant driver of gene expression variation in single cells (*43*) but also manifests at other molecular layers, including copy-number states and DNA methylation (*9*). Also, DNA replication dynamics need to be taken into consideration during experimental design and data analysis.

A starting point for many analyses can be tests for differential epigenetic profiles between different cell clusters—for example, to identify differentially methylated regions between cell types or states (*16*). In cell populations without strong substructure, it may be advantageous to quantify the epigenetic diversity of individual loci with the pairwise distance of global methylome (*16*) or estimates of epigenetic variability between cells at individual loci (*15*).

As multi-omics protocols become more widely accessible, there are also exciting opportunities to interrogate associations between different epigenetic layers and to examine associations with the transcriptome. This allows the strength of coupling between different regulatory layers to be probed in great detail. Variation in coupling strength—for example, between DNA methylation and transcription—is known from bulk analyses, comparing pluripotent to somatic cell types (*44*).

However, the variation in coupling strength can be investigated with single-cell techniques for classes of loci or individual loci between cells or between different loci within the same cell. Such variation has already been identified at different levels, including individual loci such as gene promoters and enhancers with epigenetic variation associated with expression levels of individual genes, as well as global genome-wide couplings between different layers (*22*). If multi-omics methods are applied to hybrids or outbred individuals, it may be possible to assess allele-specific methylation and expression, thereby aligning regulatory differences across molecular layers (*23*). For other analyses, it remains an open question how to best integrate data across different molecular layers. Tying together different data modalities will improve cell clustering, and the use of epigenetic information in tandem with transcriptional data will aid in reconstructing pseudotemporal orderings of cells (Fig. 4).

Adding a temporal dimension in single-cell studies

Putting multidimensional information together for each single cell gives insights not only into cell identity and function but also, through the use of different layers of the epigenome, into past history and future potential (Fig. 4A). Imagine that an otherwise stable DNA methylation mark (for example, in an imprinted gene) has changed at a specific developmental time point, which can be recorded through lineage tracing by CRISPR scarring (*45*–*49*) (Fig. 4B). This is an example in which past history is recorded. Conversely, characteristic DNA methylation patterns in induced pluripotent stem cells (iPSCs) can be predictive of the differentiation potential of these iPSCs (*50*), an example of an epigenetic state revealing future potential.

Different epigenetic marks have different stabilities in time, providing the potential to record various biological time scales. An extracellular signal acting via intracellular signaling pathways will affect transcription factor binding and thus transcription. Because transcription factor binding can be highly dynamic and nonprocessed transcripts are usually short-lived, such signals may reflect the shortest possible biological response time scale. Conversely, though, some transcription factors may bind throughout cell division (*51*) and transmit epigenetic information to the next cell generation. Different binding time scales and their functional consequences may be revealed by coupling the analysis to cell cycle state through the transcriptome (*43*). Similarly, nucleosome accessibility in promoters (or other regulatory sequences) may occur before the chromatin opening up (as may be the case with pioneer factors) or, more conventionally, allow access to transcription factors. Within one cell cycle, therefore, we can reconstruct a signaling response at its cognate promoter, giving rise to transcriptional initiation followed by the processed transcript in the cytoplasm. We can discover multiple genomic dimensions in which this signaling response plays out within this single cell. It is currently possible to reconstruct such multidimensional responses in highly synchronized tissue culture systems but not in the natural setting in vivo, let alone in complex disease situations.

The applications with the most fundamental potential for breakthroughs will also consider epigenetic memory in the system. Some epigenetic marks are heritable across cell divisions (more so in somatic cells than in early embryos), including 5mC DNA methylation, where the inheritance is very stable with a well-understood mechanism. Others, such as H3K27me3 and H3K9me2/me3, may also be inherited, although perhaps with less stability and less fidelity. Whether histone marks associated with transcriptional activation could also be heritable is an open question. A key question here is to what extent epigenetic marks are instructive (e.g., imprinting) or follow transcriptional activation or repression to lock in stabilization of cell fate decisions.

Lineage marking via single-cell sequencing methods will allow us to follow the timing of particular epigenetic changes with regard to the states before the initiation of, during, and post transcription. Furthermore, hairpin bisulphite sequencing (*52*, *53*) (in which methylation information is obtained from both DNA strands) in single cells will identify how heritable methylation is at individual loci and how heterogeneous or homogeneous such heritability is within a cell population. Measurements of 5hmC, 5fC, and 5caC across cell populations, together with mechanistic modeling approaches (*54*, *55*), will allow insights into the generation of epigenetic heterogeneity versus stable inheritance in early development, aging, and disease. The exciting prospect of single-cell epigenome editing (*39*) suggests that detailed

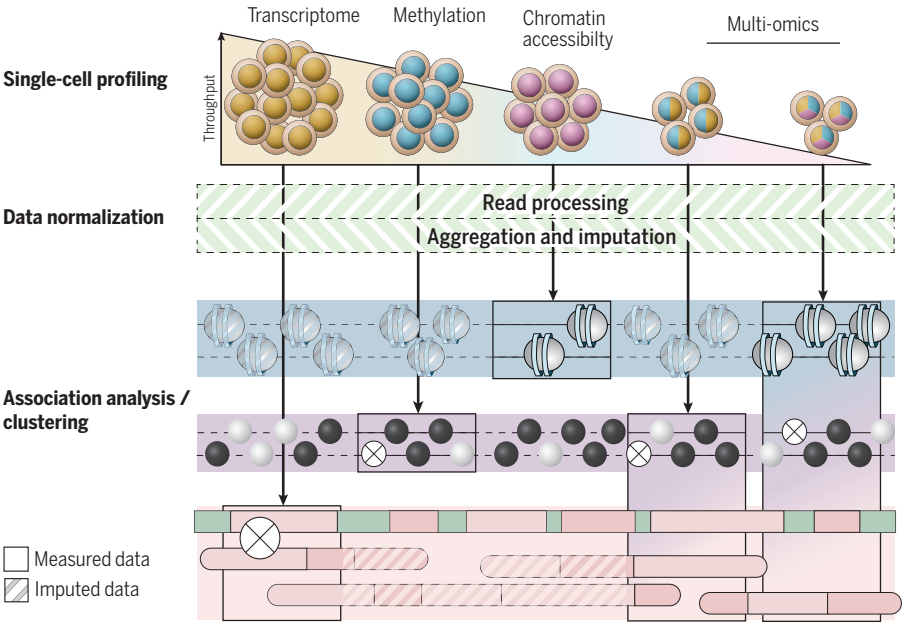

Fig. 3. Multi-omics and computational methods. Shown are typical trade-offs between single-cell RNA-seq, single-cell epigenome protocols, and multi-omics methods that provide readouts from multiple molecular layers in parallel. Consequently, it is commonly required to integrate data from different sequencing protocols. Raw sequence reads from these methods are deduplicated and aggregated into locus-specific readouts, with an optional imputation step to complete missing information. Associations between molecular layers can be used for completing missing data and allow for discovering regulatory associations.

functional testing of epigenetic marks in their various roles may soon become a reality too.

Epigenetic information may also be used to measure cell lineages (Fig. 4B). Lineage-tracing methods using CRISPR scarring have been devised (*45*–*49*), but it is not clear how accurately and reliably they work in different biological settings. Thus, DNA modifications may allow us to trace lineages by marking a particular chromosome or DNA strand, which is segregated into a particular cell type (*17*). This will be especially useful for DNA modifications that are not normally heritable (such as 5hmC, 5fC, or 5caC).

Some heritable epigenetic marks may be functionally neutral—i.e., set up in early development but simply mechanically copied at each cell division. Because the maintenance methylation machinery has a finite error rate [1 in 25 cell divisions per CpG, although this has only been measured in certain contexts (*56*)], every cell may harbor a unique code of methylation sites that would allow tracking of its developmental trajectory. This acts as if lineage were marked by DNA mutations (either natural ones or induced) (Fig. 4B). This may allow noninvasive lineaging in the future without genetic manipulation, which might be particularly useful in human studies.

We have highlighted the different time scales of variation of these different layers of the epigenome, as well as their interdependencies. It is important to recognize that most of these are from indirect measurements or inferences. In due course, we may connect epigenome dimensions by pseudotime measurements, allowing us to formulate temporal connections and dependencies. However, what is yet to materialize are real-time in vivo recording systems of epigenetic states, ideally at a single-locus level. Hence the single-cell epigenomics revolution has additional challenges to overcome. Our existing methods are already allowing us to zoom in on new concepts of "cell fate"—for example, in developmental systems where cell history can be recorded in epigenetic marks. Yet their actions at key decision points require yet unknown mechanisms (*57*, *58*). This presumably requires new epigenomic codes for cell plasticity and future potential. Deeper insights into these rules will provide not only a better understanding of living biological systems but also new tools and new ways of thinking about changing cell fate experimentally.

At the other end of the spectrum, we anticipate information regarding the presumed degradation of cell fate during aging. Models involve either clonal competition or exhaustion and hence a potential loss of cell heterogeneity in an aging tissue. Conversely, an increase in heterogeneity may occur with a concomitant loss of coherence of transcriptional networks (*59*). Interestingly, programmed changes of the epigenome during aging, particularly of the DNA methylome, accurately record chronological age. However, this "methylation aging clock" can be accelerated or decelerated by biological interventions that shorten or lengthen life span, respectively (*60*–*62*). It remains to be seen how this methylation clock plays out at the single-cell level. As many human adult diseases, including cancer, are associated with altered epigenome patterns, individual cells may gradually and in a potentially programmed way acquire disease risk via changes in epigenetic marks during aging. Conversely, single-cell multi-omics methods may identify hidden cell states with potential for tissue repair or rejuvenation.

As large-scale efforts are mapping all human cells transcriptionally and spatially [e.g., the

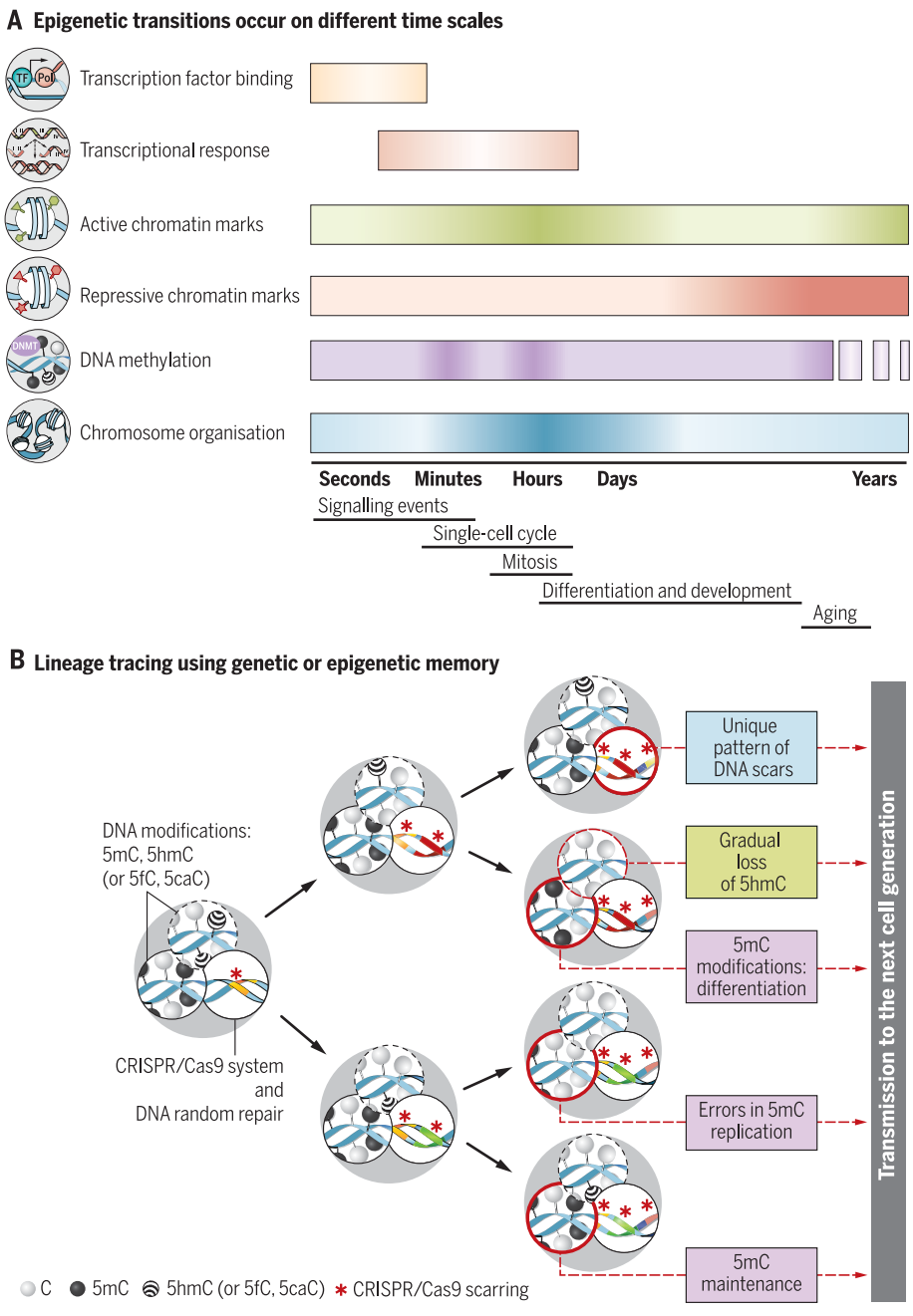

Fig. 4. Time scales of epigenetic heterogeneity at different layers and lineage tracing. (**A**) Shown are different layers of information that can be recorded at least in principle by single-cell multi-omics, from transcription factor binding and transcriptional responses to long-term epigenetic memory such as is possible with DNA methylation. Rough time scales are indicated by colored bars—with shading indicating transitions in information—and may range from seconds to years. With aging, fidelity of epigenetic information such as DNA methylation may degrade, leading to increased cell-to-cell heterogeneity. (**B**) Lineage tracing using genetic or epigenetic memory. Cell lineage can be traced by CRISPR scarring approaches in which each cell and its descendants within a lineage are linked by unique mutations or barcodes. DNA modifications may also be used to track lineage based on their inheritance and on errors in their maintenance at DNA replication. Nonheritable modifications (5hmC, 5fC, and 5caC) have a short-term lineaging potential, whereas heritable modifications (5mC) have long-term noninvasive lineaging potential.

Human Cell Atlas (*63*)], there is the prospect in the future that epigenomics measurements, in particular, will add unprecedented layers of information about memory of past experiences and about future potential of cells in the human body.

Outlook

Imagine that we had at our disposal the techniques for single-cell multi-omics, including the ability to identify all key epigenetic modalities, robustly and at an affordable cost. Imagine similarly that we had the computational tools to unravel and visualize connections between the different molecular layers within and between cells. From such advances, we anticipate answering many questions in embryonic development (including comparisons of various organisms). We would like to know any epigenetic determinants of cell fate and lineage decisions and their timing and/or memory of such decisions.

Travelling back in time (i.e., generating iPSCs) or across tissues (via transdifferentiation), we will be able to see how each cell responds in terms of erasing epigenetic memory and acquiring new cell fate trajectories, especially those not part of the normal developmental repertoire. We also anticipate unraveling tissue-level heterogeneity. Highly multiplexed methylome sequencing can already identify cell types in a complex tissue such as the brain with similar accuracy as transcriptome sequencing (*27*).

Finally, we aim to discover links between epigenetic and genetic heterogeneity, showing to what extent epigenetic change (particularly in disease) is driven by underlying changes in DNA sequence such as copy-number variation, mutations, and rearrangements in cancer, or the mobility of selfish DNA elements. Conversely, primary epimutations may underlie the initiation of some diseases but may subsequently elicit more permanent genetic change that stabilizes the disease phenotype.

These advances have implications for diagnosing and understanding disease progression. We envision that precancerous cell states may be detected at an early stage in tissues by their single-cell epigenome signatures, and other chronic diseases may also reveal unique signatures of progression. Single-cell epigenomic analyses might allow for a biopsy of only a few cells or by capturing small amounts of cell-free DNA in circulation. Such tools may also reveal cell populations in tissues with the greatest potential for regeneration and tissue repair.

REFERENCES AND NOTES

1. R. Hooke, *Micrographia: Or Some Physiological Descriptions of Minute Bodies Made by Magnifying Glasses, with Observations and Inquiries Thereupon* (Courier Corporation, 2003).
2. M. J. T. Stubbington, O. Rozenblatt-Rosen, A. Regev, S. A. Teichmann, *Science* **358**, 58–63 (2017).
3. E. Lein, L. E. Borm, S. Linnarsson, *Science* **358**, 64–69 (2017).
4. C. Gawad, W. Koh, S. R. Quake, *Nat. Rev. Genet.* **17**, 175–188 (2016).
5. C. D. Allis, T. Jenuwein, D. Reinberg, *Epigenetics* (CSHL Press, 2007).
6. W. Jin *et al.*, *Nature* **528**, 142–146 (2015).
7. J. D. Buenrostro *et al.*, *Nature* **523**, 486–490 (2015).
8. D. A. Cusanovich *et al.*, *Science* **348**, 910–914 (2015).
9. F. Guo *et al.*, *Cell Res.* **27**, 967–988 (2017).
10. S. Pott, *eLife* **6**, e23203 (2017).
11. S. J. Clark *et al.*, *bioRxiv* 138685 [Preprint] (17 May 2017).
12. A. Rotem *et al.*, *Nat. Biotechnol.* **33**, 1165–1172 (2015).
13. T. Nagano *et al.*, *Nature* **502**, 59–64 (2013).
14. H. Guo *et al.*, *Genome Res.* **23**, 2126–2135 (2013).
15. S. A. Smallwood *et al.*, *Nat. Methods* **11**, 817–820 (2014).
16. M. Farlik *et al.*, *Cell Reports* **10**, 1386–1397 (2015).
17. D. Mooijman, S. S. Dey, J. C. Boisset, N. Crosetto, A. van Oudenaarden, *Nat. Biotechnol.* **34**, 852–856 (2016).
18. C. Zhu *et al.*, *Cell Stem Cell* **20**, 720–731.e5 (2017).
19. I. C. Macaulay, C. P. Ponting, T. Voet, *Trends Genet.* **33**, 155–168 (2017).
20. I. C. Macaulay *et al.*, *Nat. Methods* **12**, 519–522 (2015).
21. S. S. Dey, L. Kester, B. Spanjaard, M. Bienko, A. van Oudenaarden, *Nat. Biotechnol.* **33**, 285–289 (2015).
22. C. Angermueller *et al.*, *Nat. Methods* **13**, 229–232 (2016).
23. Y. Hu *et al.*, *Genome Biol.* **17**, 88 (2016).
24. Y. Hou *et al.*, *Cell Res.* **26**, 304–319 (2016).
25. C. Angermueller, H. J. Lee, W. Reik, O. Stegle, *Genome Biol.* **18**, 67 (2017).
26. M. Frommer *et al.*, *Proc. Natl. Acad. Sci. U.S.A.* **89**, 1827–1831 (1992).
27. C. Luo *et al.*, *Science* **357**, 600–604 (2017).
28. R. M. Mulqueen *et al.*, *bioRxiv* 157230 [Preprint] (2 June 2017).
29. L. F. Cheow, S. R. Quake, W. F. Burkholder, D. M. Messerschmidt, *Nat. Protoc.* **10**, 619–631 (2015).
30. J. R. Peat *et al.*, *Cell Reports* **9**, 1990–2000 (2014).
31. X. Wu, A. Inoue, T. Suzuki, Y. Zhang, *Genes Dev.* **31**, 511–523 (2017).
32. M. R. Corces *et al.*, *Nat. Genet.* **48**, 1193–1203 (2016).
33. S. Preissl *et al.*, *bioRxiv* 159137 [Preprint] (6 July 2017).
34. J. Kind *et al.*, *Cell* **163**, 134–147 (2015).
35. I. M. Flyamer *et al.*, *Nature* **544**, 110–114 (2017).
36. T. Nagano *et al.*, *Nature* **547**, 61–67 (2017).
37. V. Ramani *et al.*, *Nat. Methods* **14**, 263–266 (2017).
38. T. J. Stevens *et al.*, *Nature* **544**, 59–64 (2017).
39. J. van Arensbergen, B. van Steensel, *Mol. Cell* **66**, 167–168 (2017).
40. S. Gravina, X. Dong, B. Yu, J. Vijg, *Genome Biol.* **17**, 150 (2016).
41. W. Zhang, T. D. Spector, P. Deloukas, J. T. Bell, B. E. Engelhardt, *Genome Biol.* **16**, 14 (2015).
42. J. Ernst, M. Kellis, *Nat. Biotechnol.* **33**, 364–376 (2015).
43. F. Buettner *et al.*, *Nat. Biotechnol.* **33**, 155–160 (2015).
44. G. Ficz *et al.*, *Cell Stem Cell* **13**, 351–359 (2013).
45. A. McKenna *et al.*, *Science* **353**, aaf7907 (2016).
46. J. P. Junker *et al.*, *bioRxiv* 056499 [Preprint] (1 June 2016).
47. S. D. Perli, C. H. Cui, T. K. Lu, *Science* **353**, aag0511 (2016).
48. R. Kalhor, P. Mali, G. M. Church, *Nat. Methods* **14**, 195–200 (2017).
49. K. L. Frieda *et al.*, *Nature* **541**, 107–111 (2017).
50. M. Nishizawa *et al.*, *Cell Stem Cell* **19**, 341–354 (2016).
51. X. Huang, J. Wang, *Cell Stem Cell* **20**, 741–742 (2017).
52. C. D. Laird *et al.*, *Proc. Natl. Acad. Sci. U.S.A.* **101**, 204–209 (2004).
53. L. Zhao *et al.*, *Genome Res.* **24**, 1296–1307 (2014).
54. F. von Meyenn *et al.*, *Mol. Cell* **62**, 983 (2016).
55. P. Giehr, C. Kyriakopoulos, G. Ficz, V. Wolf, J. Walter, *PLOS Comput. Biol.* **12**, e1004905 (2016).
56. T. Ushijima *et al.*, *Genome Res.* **13**, 868–874 (2003).
57. H. J. Lee, T. A. Hore, W. Reik, *Cell Stem Cell* **14**, 710–719 (2014).
58. H. Mohammed *et al.*, *Cell Reports* **20**, 1215–1228 (2017).
59. C. P. Martinez-Jimenez *et al.*, *Science* **355**, 1433–1436 (2017).
60. S. Horvath, *Genome Biol.* **14**, R115 (2013).
61. G. Hannum *et al.*, *Mol. Cell* **49**, 359–367 (2013).
62. T. M. Stubbs *et al.*, *Genome Biol.* **18**, 68 (2017).
63. A. Regev *et al.*, *bioRxiv* 121202 [Preprint] (8 May 2017).
64. B. Lake *et al.*, *bioRxiv* 128520 [Preprint] (19 April 2017).

ACKNOWLEDGMENTS

W.R. thanks I. Herraez, T. Stubbs, S. Clark, C. Alda, H. Mohammed, M. Eckersley-Maslin, S. Rulands, W. Dean, J. Marioni, and B. Simons for discussions or comments on the manuscript. Thank you to V. Juvin (SciArtWork) for artwork. Work in W.R.'s laboratory is supported by the Wellcome Trust, the Biotechnology and Biological Sciences Research Council (BBSRC), and the Medical Research Council (MRC): G.K. is supported by the BBSRC and the MRC; O.S. is supported by European Molecular Biology Laboratory core funding, the Wellcome Trust, and the European Research Council. W.R. is a consultant and shareholder of Cambridge Epigenetix.

10.1126/science.aan6826

Single-cell epigenomics: Recording the past and predicting the future

Gavin Kelsey, Oliver Stegle and Wolf Reik

Science **358** (6359), 69-75.
DOI: 10.1126/science.aan6826