nature biotechnology

Nondestructive, base-resolution sequencing of 5-hydroxymethylcytosine using a DNA deaminase

Emily K Schutsky[1], Jamie E DeNizio[1], Peng Hu[2], Monica Yun Liu[1], Christopher S Nabel[1], Emily B Fabyanic[2], Young Hwang[3], Frederic D Bushman[3], Hao Wu[2,4] & Rahul M Kohli[1,4]

Here we present APOBEC-coupled epigenetic sequencing (ACE-seq), a bisulfite-free method for localizing 5-hydroxymethylcytosine (5hmC) at single-base resolution with low DNA input. The method builds on the observation that AID/APOBEC family DNA deaminase enzymes can potently discriminate between cytosine modification states and exploits the non-destructive nature of enzymatic, rather than chemical, deamination. ACE-seq yielded high-confidence 5hmC profiles with at least 1,000-fold less DNA input than conventional methods. Applying ACE-seq to generate a base-resolution map of 5hmC in tissue-derived cortical excitatory neurons, we found that 5hmC was almost entirely confined to CG dinucleotides. The whole-genome map permitted cytosine, 5-methylcytosine (5mC) and 5hmC to be parsed and revealed genomic features that diverged from global patterns, including enhancers and imprinting control regions with high and low 5hmC/5mC ratios, respectively. Enzymatic deamination overcomes many challenges posed by bisulfite-based methods, thus expanding the scope of epigenome profiling to include scarce samples and opening new lines of inquiry regarding the role of cytosine modifications in genome biology.

Epigenetic modification of cytosine bases is crucial for proper regulation of gene expression in mammals[1]. Although 5mC is best characterized for its gene-repressive roles, the types of known modifications greatly expanded with the identification of several oxidized forms of 5mC (ox-mCs) that arise via the action of ten-eleven translocation (TET) family enzymes[2–5]. Ox-mCs serve as intermediates in active DNA demethylation, whereby repressive 5mC marks are erased; ox-mCs may also have independent epigenetic functions[6]. 5hmC is by far the most abundant ox-mC, reaching levels as high as 1.8% of total cytosines in human neurons, where 5mC comprises 4–5% of total cytosines[7]. The highly oxidized bases 5-formylcytosine (5fC) and 5-carboxylcytosine (5caC) have also been detected, but are far less common: when quantified in parallel with 5hmC, 5fC was maximally detected at levels greater than three orders of magnitude less (0.0007% of total cytosines in human neurons)[7], whereas 5caC quantified around 5 ppm of total cytosine in mouse embryonic stem cells (mESCs) and was undetectable in neurons[3,8].

The most commonly used approaches for localizing cytosine modifications rely on differential chemical reactivity of cytosine variants in bisulfite sequencing (BS-seq; **Fig. 1a**)[9]. Incubation of DNA with bisulfite at extreme pH and elevated temperature promotes deamination of cytosine to uracil, whereas 5mC is largely unreacted (**Supplementary Fig. 1**). With the discovery of ox-mCs, the interpretation of BS-seq became more complicated. Although 5fC and 5caC deaminate in BS-seq, 5hmC forms a bulky adduct that is slow to deaminate, rendering 5hmC indistinguishable from 5mC[10]. To localize 5hmC specifically, several techniques have been advanced to change the bisulfite reactivities of 5mC versus 5hmC. In TET-assisted bisulfite sequencing (TAB-seq)[11], 5hmC bases are enzymatically modified by glucosylation (yielding 5ghmC), and 5mC is then selectively oxidized to 5caC *in vitro* by TET. Following bisulfite treatment, all bases except the protected 5ghmC are deaminated. Alternatively, oxidative bisulfite sequencing (oxBS-seq) employs selective oxidization of 5hmC to 5fC before bisulfite conversion[12]. Subtraction of oxBS-seq from standard BS-seq signals allows for indirect identification of 5hmC.

A key limitation of these methods is the use of bisulfite, as chemical deamination conditions can degrade as much as 99.9% of input genomic DNA (gDNA)[13]. Thus, when samples of gDNA are limiting, experiments either provide limited coverage of the genome[14], or methods using reduced representation or enrichment steps need to be performed[15]. Thus, the characterization of specific primary cell types, as well as rare cell populations undergoing dynamic epigenetic changes, has been challenging, emphasizing the importance of methods that permit low amounts of starting DNA. For 5hmC detection, methods that avoid the use of bisulfite have been pursued, including nanopore, single-molecule real time (SMRT), and restriction-enzyme-based approaches[16–19]. However, the limitations of each method have contributed to conflicting results, leaving, for example, the prevalence of 5hmC in non-CG sites a matter of debate[20–23].

Recognizing the constraints of chemical deamination, we were drawn toward a natural analog of this reaction: enzymatic deamination by an AID/APOBEC family DNA deaminase. Here we report the development, validation and application of ACE-seq. In our approach, deamination under near-physiological, non-destructive conditions permitted single-base resolution 5hmC profiling with minimal DNA input.

[1]Department of Medicine, University of Pennsylvania, Philadelphia, Pennsylvania, USA. [2]Department of Genetics, University of Pennsylvania, Philadelphia, Pennsylvania, USA. [3]Department of Microbiology, University of Pennsylvania, Philadelphia, Pennsylvania, USA. [4]Penn Epigenetics Institute, University of Pennsylvania, Philadelphia, Pennsylvania, USA. Correspondence should be addressed to R.M.K. (rkohli@pennmedicine.upenn.edu) or H.W. (haowu2@pennmedicine.upenn.edu).

Received 6 February; accepted 3 July; published online 8 October 2018; doi:10.1038/nbt.4204

RESULTS

Development of ACE-seq

Members of the AID/APOBEC family catalyze the deamination of cytosine to uracil in single-stranded DNA (ssDNA) and mediate critical functions in innate and adaptive immunity[24]. Previously, we found that several family members can discriminate between cytosine modification states[25]; however, their overall poor catalytic activity prevented biotechnological applications. More recently, motivated by subsequent studies demonstrating that one human-specific family member, APOBEC3A (A3A), has high activity and a particular proficiency for 5mC deamination[26,27], our attention turned to quantifying A3A's activity on the full spectrum of natural cytosine modifications. We found that A3A indeed readily deaminated C and 5mC, but also discriminated potently against all three ox-mCs, with a ~5,000-fold reduction in the 5hmC deamination rate relative to that of cytosine[28]. This observation raised the prospect that A3A's differential reactivity could be exploited to localize 5hmC in gDNA without the need for bisulfite.

We envisioned three potential barriers to using AID/APOBECs in sequencing. First, enzymatic deamination, like chemical deamination, requires ssDNA; unlike in BS-seq, gDNA would need to be denatured under mild conditions that permit enzymatic activity in ACE-seq. Second, different AID/APOBECs show distinctive sequence preferences, with 5′ bases influencing activity most strongly. A3A, for example, preferentially deaminates TTC, with reduced activity in disfavored contexts[22]. In ACE-seq, sequence context preferences would have to be overcome. Third, ACE-seq would require a sufficient window of enzymatic selectivity such that false-positive (C/5mC non-conversion) and false-negative (5hmC conversion) readouts would be minimized.

As a model system to develop ACE-seq, we used phage gDNA, as these genomes offer known, homogenous modifications (**Supplementary Fig. 2**). The 169-kb T4 phage genome normally contains all cytosine bases replaced by 5ghmC (here referred to as T4-ghmC). Established mutants in the 5ghmC pathway yield phage in which every encoded cytosine is 5hmC (referred to as T4-hmC) or in which all cytosine bases are unmodified (referred to as T4-C)[29]. To first evaluate conditions for ssDNA generation, we subjected 1 ng of gDNA from T4-C to brief heat denaturation, snap freezing, and subsequent incubation with excess A3A (5 μM). A target locus was then amplified and cloned, and individual clones were sequenced to guide method development. Some secondary structure elements proved to be resistant to deamination, but this was overcome by denaturing in the presence of DMSO and performing A3A incubation under ramping temperature conditions (**Supplementary Fig. 3**). Efficient and complete deamination of the T4-C target locus was observed, as determined by sequencing of five individual clones (**Fig. 2a**).

When these conditions were next applied to T4-hmC, the majority of 5hmC bases were protected from deamination (**Fig. 2b**); however, some 5hmC deamination events were observed (~10% of cytosines across three separate clones). These events aligned with A3A preferences (that is, TThmC sites), validating the prior observation that 5hmC can be deaminated by excess A3A, albeit inefficiently[28]. Having previously shown that AID/APOBECs discriminate against bulky 5-position modifications[25,28], we considered whether further modification could protect 5hmC from deamination. Indeed, under conditions in which some T4-hmC deamination occurs, we observed 0% deamination of T4-ghmC at any position across four clones sequenced (**Fig. 2c**). We therefore posited that *in vitro* glucosylation of 5hmC in gDNA could be used to prevent its sporadic deamination by A3A.

To simultaneously examine cytosine, 5mC and 5hmC deamination, we pooled lambda (λ) phage gDNA (48.5 kb) that was enzymatically methylated at all CG sites (**Supplementary Fig. 2**) with T4-hmC. The pooled gDNA was sheared, and a low input sample (1 ng) was treated with T4 β-glucosyltransferase (βGT) to convert 5hmC in T4-hmC to 5ghmC, followed by incubation with excess A3A. The resulting products were analyzed using Illumina high-throughput sequencing after library preparation. Unbiased analysis of the λ phage genome showed robust cytosine/5mC deamination: non-conversion of cytosine or 5mCG was detected for only ~6,800 of 2.4 million independently sequenced cytosine sites and ~10,500 of ~800,000 5mCG sites, respectively (**Fig. 2d** and **Supplementary Table 1**). The cytosine and 5mC non-conversion rates (0.3% and 1.3%, respectively) were similar or better than those observed with TAB-seq (0.4%; 2.2%) on a comparable λ gDNA control[11]. In analyzing the enzymatically glucosylated T4-hmC phage, ~99.4% of all 5hmC bases were called as cytosine in sequencing (**Fig. 2d**), exceeding the sensitivity of TAB-seq (84.4–92.0%)[11,23,30]. We also digested the ACE-seq-treated phage gDNA to nucleosides and quantified it via liquid chromatography coupled to tandem mass spectrometry (LC-MS/MS). Efficient conversion of cytosine and 5mC (~97% decrease in each signal) and protection of 5hmC were observed (**Fig. 2e**), and the necessity of βGT protection of 5hmC was confirmed (**Supplementary Fig. 4**). Taken together, both sequencing- and LC-MS/MS-based approaches orthogonally confirmed the robust conversion and protection efficiencies of ACE-seq.

Thus, regarding the three barriers initially considered, in the optimized ACE-seq protocol (**Fig. 1b**), a modified denaturation step permitted A3A to access its ssDNA substrate and driving conditions with excess A3A allowed full C/5mC deamination in all sequence contexts. Although these conditions resulted in some 5hmC deamination, these bases could be fully protected from deamination by glucosylation to generate a wide window for cytosine/5mC versus 5hmC discrimination.

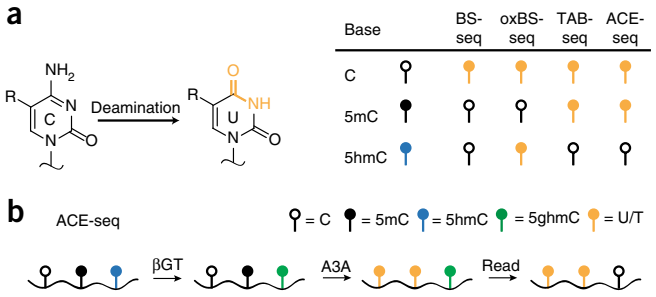


Figure 1 Reactivities of modified cytosines in sequencing approaches. (**a**) Deamination underlies the differentiation of modified cytosines in current sequencing approaches. Standard BS-seq converts cytosine to uracil and leaves 5mC and 5hmC unconverted, reading as cytosine in sequencing. Modifications to 5mC and 5hmC in oxBS-seq and TAB-seq, when coupled to bisulfite, can differentiate between these two bases. ACE-seq uses enzymatic rather than bisulfite-mediated deamination to provide a readout that is comparable to that of TAB-seq. (**b**) In the optimized ACE-seq protocol, APOBEC3A catalyzes the enzymatic deamination of C and 5mC to completion, whereas 5hmC, which is highly resistant to deamination, but is further protected by glucosylation, is localized by its non-conversion.

ACE-seq is non-destructive

In typical bisulfite-based approaches, template gDNA damage limits the size of amplicons that can be characterized to those typically <300 bp[31]. For direct comparison, we treated mESC gDNA using either BS-seq or ACE-seq conditions. After deamination, we attempted to amplify either short (200 bp) or long (1 kb) amplicons from a single

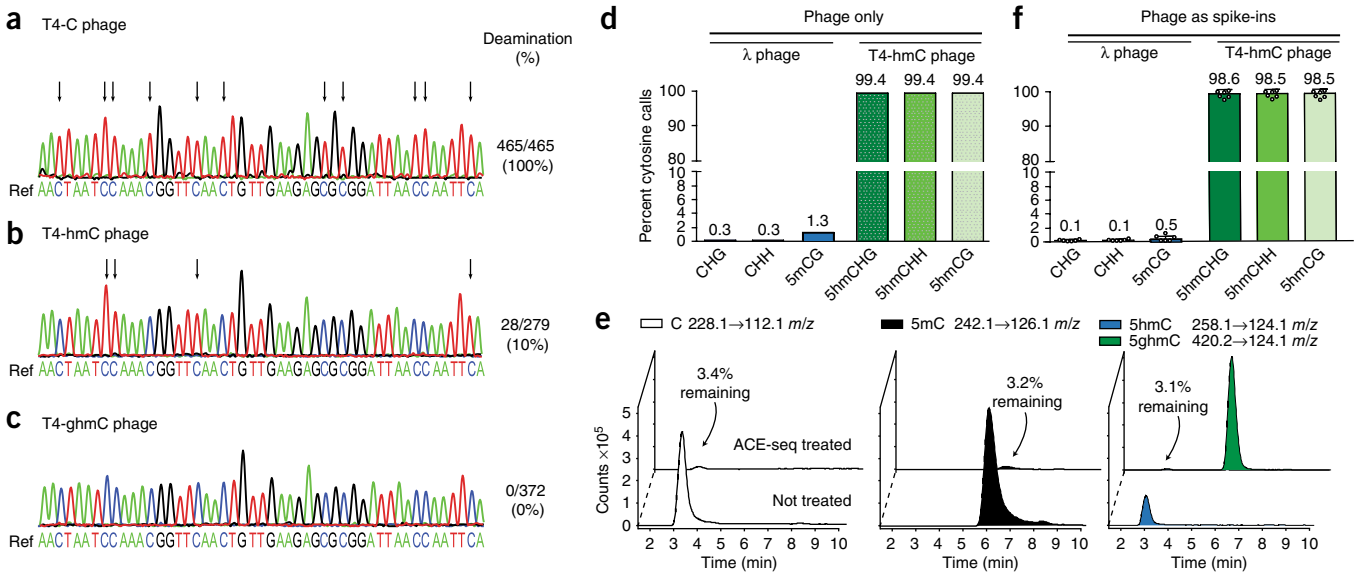


Figure 2 Development and validation of ACE-seq. (**a**–**c**) 1-ng samples of T4 phage gDNA with homogeneous modifications (**a**, T4-C; **b**, T4-hmC; **c**, T4-ghmC) were heated, snap frozen and incubated with A3A before amplification of a genomic segment, TA cloning and Sanger sequencing of individual clones. Illustrative sequencing traces from individual clones are shown above the reference genome. Arrows denote deamination events (C>T transitions). Deamination events are quantified as the number of cytosines that were deaminated across the sum of all clones (93 cytosines per clone; T4-C, five clones; T4-hmC, three clones; T4-ghmC, four clones). (**d**,**f**) Rates of non-conversion for enzymatically methylated λ phage gDNA (5mCG, CH) and T4-hmC phage gDNA in ACE-seq, as determined by Illumina sequencing, using inputs of either 1 ng of each alone (**d**) or 100 pg each (**f**) as spike-ins averaged across six mammalian DNA samples (see **Supplementary Table 1**). Mean values are listed above each bar, and error bars represent s.d. Individual data points are overlaid on the plot. (**e**) Representative LC-MS/MS traces of cytosine, 5mC, 5hmC and 5ghmC nucleosides after a 1:1 mix of methylated λ gDNA and T4-hmC gDNA was subjected to ACE-seq treatment (compared with untreated control sample). Percentages denote the amount detected after ACE-seq treatment, averaged across three independent replicates.

target locus (*Tbx5*) using a fixed number of PCR cycles. For the short amplicon, although the ACE-seq sample was more readily amplified, amplicons could be detected below 10 ng of input gDNA in either condition (**Fig. 3a**). However, with the 1-kb locus, ACE-seq amplicons were detectable with ~3-log lower DNA input than with BS-seq, suggesting that BS-seq introduced substantial damage in the template (**Fig. 3b**). Notably, the 1-kb amplicons with ACE-seq amplified nearly as efficiently as the 200-bp amplicons, suggesting that the gDNA stayed intact under ACE-seq conditions. We validated this finding using quantitative PCR (**Fig. 3c** and **Supplementary Fig. 5**), in which BS-seq shifted the threshold cycle number by >6.0 relative to ACE-seq, suggesting a >64-fold decrease in intact template for the 1-kb amplicon.

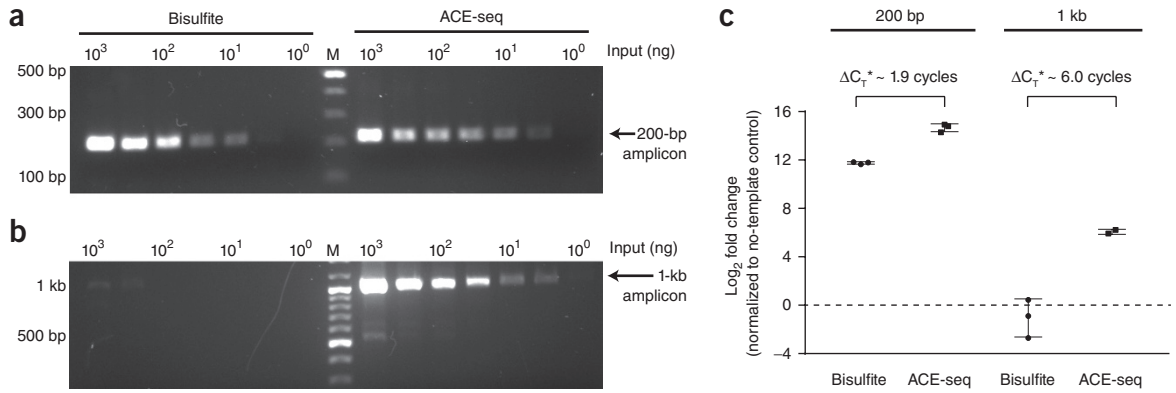


Figure 3 ACE-seq is non-destructive. Initial input levels of gDNA from mESCs were titrated from 1 μg to 1 ng and the samples were treated with either BS-seq or ACE-seq protocols. (**a**,**b**) Primers were designed to amplify either a 200-bp amplicon (**a**) or a 1-kb amplicon (**b**) from the *Tbx5* genomic locus, using 35 cycles of PCR. Resulting amplicons were run on 1.5% agarose gels and stained with SybrSafe. Marker (M) is in the middle lane with bold bands at 1 kb and 500 bp. The bisulfite experiment was performed twice with similar results and used to inform conditions for the ACE-seq experiment. (**c**) The samples were also analyzed by quantitative PCR (qPCR). The difference between threshold cycle (C_T) in the absence of template (water-only control) versus reactions containing 1 μg of template is plotted. In **a** and **b**, the input amount was normalized, whereas the volume input in **c** was normalized with qPCR, resulting in twofold less BS-seq input relative to ACE-seq. Data are reported as collected, but ΔC_T* denotes subtraction of 1 cycle from ACE-seq measurements to account for a twofold difference in initial input resulting from dilution. qPCR data for other input concentrations of gDNA are reported in **Supplementary Figure 5**. Individual triplicate data points are plotted, and error bars represent the s.d.

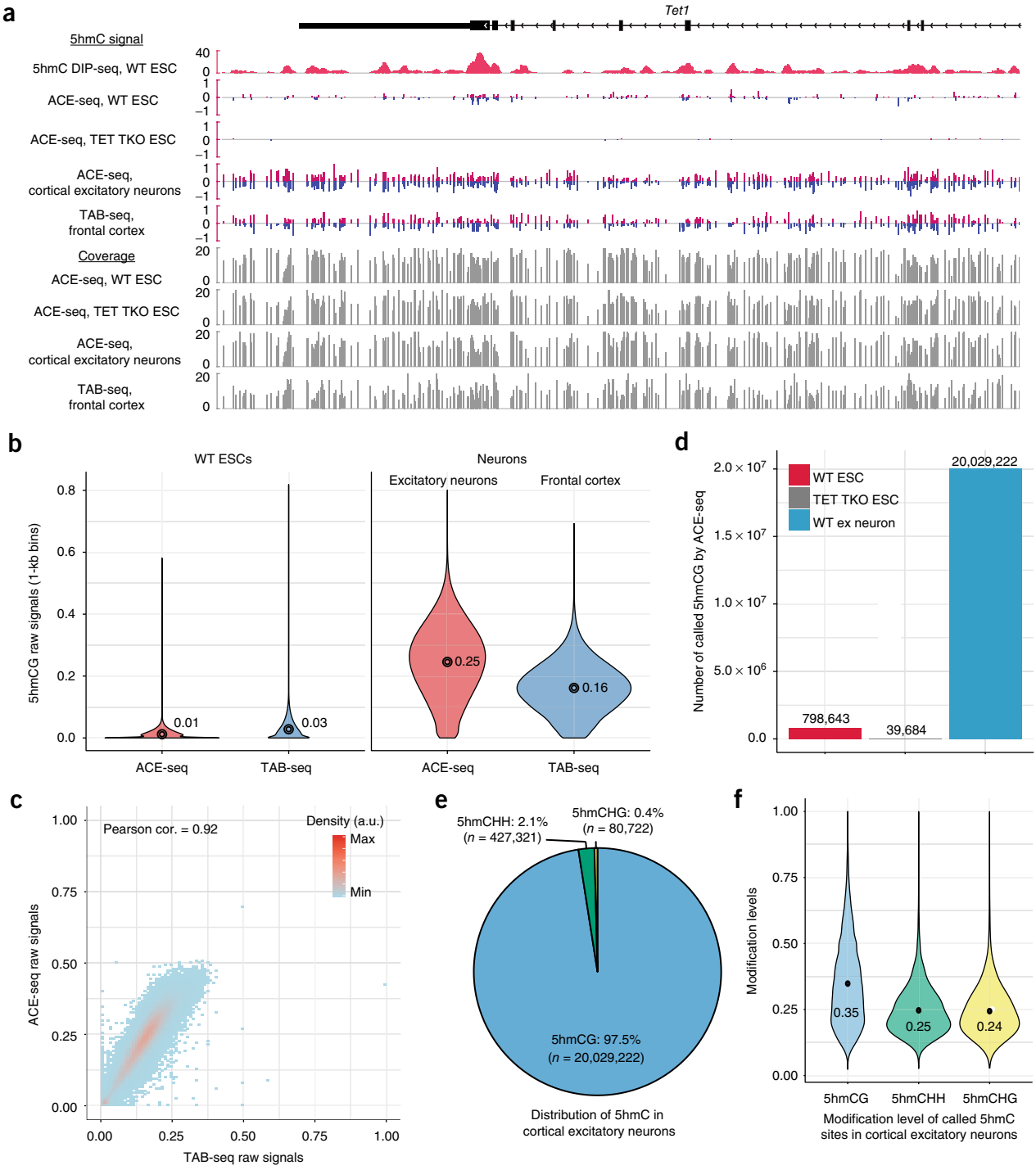


Figure 4 Generation of whole-genome base-resolution maps of 5hmC using ACE-seq. (**a**) Snapshot of base-resolution 5hmC maps (ACE-seq or TAB-seq: red/blue for positive (Watson)/negative (Crick) strands, respectively) compared with DNA-immunoprecipitation-based 5hmC map (DIP-seq: pink) near the *Tet1* locus (chr10:62,262,357–62,300,848; genome build: mm9). Only CGs covered by at least two reads are shown. Gray tracks denote sequencing coverage. All ACE-seq data shown represent merged data sets from single experiments with 2 ng and 20 ng of input DNA ($n = 2$). (**b**) Violin plots comparing raw 5hmCG signals (fraction of C/(C+T)) in 1-kb genomic bins between ACE-seq (red) or TAB-seq (blue), with mean values above each plot. The width of the violin plot corresponds to the kernel probability density of the data at a given value, and the circle indicates the mean value. (**c**) Correlation density plot between ACE-seq signals in neurons and TAB-seq signals in frontal cortex (in 10-kb bins). a.u., arbitrary units. Correlation analysis was performed with 10-kb bins spanning the genome ($n = 238{,}401$). (**d**) Bar graph of statistically significant 5hmCG sites (P value = 2.5×10^{-4}) in WT mESCs, TET TKO mESCs and WT cortical excitatory (ex) neurons, with values listed above each bar. (**e**) Sequence context of statistically significant 5hmC sites (in WT excitatory neurons) compared with the reference mouse genome (for 5hmCH sites, P value = 5×10^{-8}). (**f**) Violin plot of the distribution of modification levels of called 5hmCG, 5hmCHH and 5hmCHG sites in WT ex neurons, with mean values listed. The width of the violin plot corresponds to the kernel probability density of the data at a given value.

ACE-seq detects true-positive 5hmC bases

We next characterized 5hmC at single-base resolution in mammalian cells. To highlight ACE-seq's utility on a specific tissue-derived cell subtype, we isolated gDNA from purified murine cortical excitatory neurons (*NeuroD6/NEX*+)[32] and compared it with gDNA from total mouse brain cortex previously characterized by TAB-Seq[23]. gDNA was also purified from wild-type and TET-triple knockout (TKO) mESCs, the latter of which lack authentic 5hmC. For each sample, gDNA was spiked with 0.5% of T4-hmC and CG-methylated λ phage gDNA as internal controls. Either 2 or 20 ng of gDNA was then subjected to the optimized ACE-seq protocol and sequenced to an average depth of ~8-11× per strand, comparable to previous TAB-seq experiments (**Supplementary Table 1**). Reads were aligned and complete strand-specific base-resolution 5hmC maps were established for each of the three samples (**Fig. 4a**). Analysis of the spike-in controls confirmed an average non-conversion rate of 5hmC (98.5%) versus cytosine/5mC (0.1%/0.5%, respectively) (**Fig. 2f**), enabling robust 5hmC discrimination.

We first analyzed the raw 5hmC signal detectable in the CG context. 5hmCG was highly abundant in the cortical excitatory neurons (mean = 25%), but less abundant in mESCs (mean = 1%), as anticipated (**Fig. 4b**). Given the high prevalence in neurons, we performed a pairwise comparison between the 5hmC signal from ACE-seq in cortical excitatory neurons and that from TAB-seq of the brain cortex, and found a strong correlation (r = 0.92; **Fig. 4c**) consistent with the fact that the cortex contains ~85% excitatory neurons. Notably, only 2 or 20 ng of input gDNA was used in ACE-seq, as compared with ~3 μg of cortical gDNA in TAB-seq. A comparison between data collected using 2 ng or 20 ng of gDNA with ACE-seq also revealed high correlation, indicating that ACE-seq requires substantially less input DNA than bisulfite-based methods (**Supplementary Fig. 6**).

We next performed statistical calling of 5hmC sites in CGs. Using a P value cut-off of 2.5×10^{-4}, we called 798,643 high-confidence 5hmCG sites in wild-type (WT) mESCs (**Fig. 4d**). By benchmarking ACE-seq against TET TKO ESCs, our study offers the ability to empirically determine, rather than statistically estimate, the false-discovery rate (FDR) of such a sequencing approach for the first time. Using the same statistical framework applied to WT ESCs, we detected 39,684 false-positive 5hmCG sites in TET TKO ESCs, resulting in an FDR of ~5.0% in these samples, identical to that estimated in TAB-Seq[11]. In neurons, where 5hmC is especially abundant, more than 20 million 5hmC sites were readily called, resulting in a calculated FDR of ~0.2%. Although detection of 5fC and 5caC is another theoretical source of false discovery as A3A discriminates against all ox-mCs[28], the 5hmC abundance significantly exceeds that of 5fC/5caC in all cell and tissue types (>100-fold in mESCs[5]; >1,000-fold in mouse/human brain[7,8]), making the contribution of 5fC/5caC to the ACE-seq signal negligible. Thus, ACE-seq permitted high-confidence 5hmC calling in both our ESC and neuronal samples.

5hmC in non-CG contexts is rare

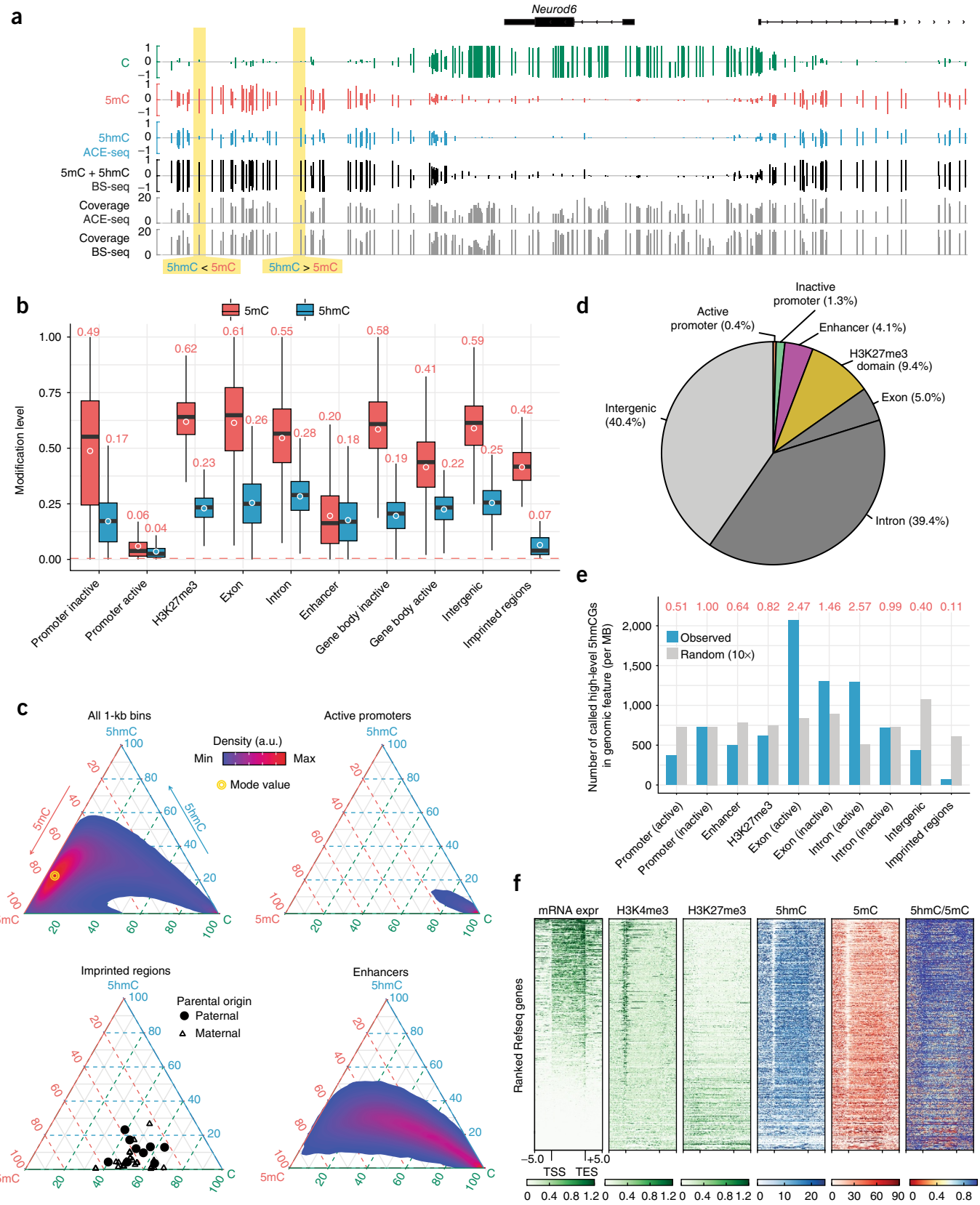
One area of disagreement in prior profiling has centered on the level and importance of 5hmC in non-CG (CH) contexts. When subtractive oxBS-seq or a less-quantitative restriction-enzyme-based method was applied to neuronal samples, 5hmCH levels were suggested to be relatively prevalent[21,22]. By contrast, TAB-seq analysis of mouse cortical DNA revealed that 5hmC occurs almost exclusively in CG contexts (1.9% of 5hmC in CH contexts), with detectable, but low, levels in ESCs (0.11% in CH)[11,23]. Using A3A provides an orthogonal approach to these methods and, unlike TET in TAB-seq, which has a preference for oxidation of CGs[33], has the added strength of being agnostic to the 3′ base[28]. Using a P value cut-off of 5×10^{-8}, the FDR for 5hmCH detection in cortical excitatory neurons was 4.0%, obtained from analyzing the sequence-context-matched TET TKO samples. Using this statistical framework, ACE-seq confirmed that 5hmC is a rare modification in CH contexts, with only ~2.5% of 5hmC as 5hmCH (5hmCHH: 427,321 sites; 5hmCHG: 80,722 sites) in cortical excitatory neurons (**Fig. 4e**). Despite their rarity, the levels of 5hmC modification at called 5hmCH sites were largely comparable to those at 5hmCG sites (**Fig. 4f**). For mESCs, the direct comparison with TET TKO matched controls allows us to state with statistical confidence that 5hmCH was not detectable in ESCs (that is, called 5hmCH in WT ESCs did not exceed that in TET TKO ESCs) at the current sequencing depth.

5hmCG and 5mCG genomic distribution in excitatory neurons

Having established the 5hmC landscape in cortical excitatory neurons, we next integrated our analysis with a neuronal-subtype-matched BS-seq data set (*Camk2a*+)[30] to uncover the true composition and genomic patterning of cytosine modifications in murine cortical excitatory neurons. By subtracting the ACE-seq signals from those of BS-seq, we constructed single-base resolution maps of cytosine, 5mC and 5hmC, revealing that sites that appeared to be fully methylated in BS-seq could vary substantially in terms of 5hmC and 5mC distribution (**Fig. 5a** and **Supplementary Fig. 7**). Across various classes of gene regulatory elements and genomic features, the levels of 5hmC were generally less than those of 5mC, and approximate 5hmC/5mC ratios remained nearly constant across many genomic features (5hmC/5mC ratio ~0.3–0.5, excluding outliers; **Fig. 5b**).

Figure 5 Genomic distribution of 5hmC and 5mC in adult cortical excitatory neurons. (**a**) Snapshot of base-resolution cytosine (green), 5mC (red) and 5hmC (blue) maps near the *Neurod6* gene locus (chr6:55,667,934–55,690,102; genome build: mm10). Only CGs covered by at least two reads are shown. Gray tracks denote sequencing coverage. All of the ACE-seq data shown represent merged data sets from single experiments with 2 ng and 20 ng of input DNA (n = 2). (**b**) The modification levels of 5hmCG (blue) and 5mCG (red) for several classes of genomic elements. The red dashed line denotes the 5mCG non-conversion rate in ACE-seq. Genic features were extracted from the UCSC Refseq Genes database and imprinted regions were chosen from a prior study[34]. Transcriptional activity of promoters/gene bodies and the presence of H3K27me3-marked repressive domains reflect experimentally determined results in *Camk2a*-positive cortical excitatory neurons from H3K4me3 and H3K27me3 ChIP-seq experiments[23]. Enhancers (>1 kb from transcriptional start site, TSS) are determined by ATAC-seq experiments in *Camk2a*-positive cortical excitatory neurons[23]. The white circles denote the mean of 5mC or 5hmC levels across the indicated genomic elements, with mean values listed above each plot. The center line represents the median value, the box represents the interquartile range, and the whiskers represent the maximum and minimum values. (**c**) Ternary plots show the levels of cytosine, 5mC and 5hmC in 1-kb bins across the genome (all) or overlapping with representative genomic elements. (**d**) Pie chart shows the overlap of called 5hmCGs with genomic elements. Each called 5hmCG site is counted once: the overlap of a genomic element excludes all previously overlapped sites clockwise starting from active promoter. (**e**) The relative enrichment of 5hmCG (blue) and random sites (gray) at genomic elements (normalized to the coverage of the element type). 'Random' consists of ten random samplings of genomic elements. Shown on the top are the ratios (red) between observed and random. (**f**) Heat map representation of normalized RNA-seq, H3K4me3 (ChIP-seq), H3K27me3 (ChIP-seq), 5hmC (ACE-seq), 5mC (derived from BS-seq and ACE-seq) and 5hmC/5mC ratios in 33,136 mouse Refseq genes (gene length > 200 bp). Genes were ranked by their expression levels in *Camk2a*-positive cortical excitatory neurons.

Our analysis revealed a few notable exceptions to the observation that 5hmC levels track with 5mC levels in most features. First, at promoter-distal enhancers (identified by ATAC-seq), 5hmC levels were nearly equivalent to those of 5mC (5hmC/5mC ratio = 0.90), rather than being proportionally low, as would be expected from the overall genomic trend (**Fig. 5b**). A compelling counterexample was



provided by imprinting control regions (ICRs), which control parent-of-origin-dependent allelic-specific expression. Whole genome BS-seq analysis previously identified differentially methylated ICRs in the mouse brain[34], but the contribution of 5hmC remained unknown. We focused on 30 differentially methylated regions (DMRs) in imprinting regions (including bona fide ICRs; **Supplementary Table 2**) and found that the average 5hmC level (6.7%) was well below the genomic baseline level, and even lower than expected relative to 5mC (5hmC/5mC ratio ~0.17; **Fig. 5b**). The majority of imprinted regions exhibited this unique 5hmC/5mC patterning, as 22 of 30 imprinted regions showed 5hmC levels < 10%, and 5hmC/5mC ratios averaging 0.08 (**Fig. 5c** and **Supplementary Fig. 8**). These observations imply that TET binding or activity is negatively regulated at most imprinted regions to maintain allelic-specific 5mCG patterns in neurons.

To explore the intra-class heterogeneity, we analyzed tiled 1-kb genomic bins for cytosine, 5mC and 5hmC distribution and generated ternary plots that accounted for all three modification states (**Fig. 5c**). Across these bins, a wide distribution of states could be observed centered on the mode value of ~6% CG, ~72% 5mCG and ~22% 5hmCG. Although some genomic features, such as genic regions including exons, tracked well with the overall distribution, others showed substantial deviations (**Supplementary Fig. 9**). For example, regions overlapping with active promoters largely exhibit a homogeneous, hypomodified state, whereas enhancers display a 'long tail' in the ternary plot, reflecting highly heterogeneous, partially modified states of 5hmC and 5mC (**Fig. 5c**). These observations are notable in the context of the recent discovery that transcription factor (TF) binding can be differentially influenced by methylation of their target sequences[35]. The heterogeneity in 5hmC/5mC that we observed may similarly influence TF activity in these regulatory regions.

Given the heterogeneity evident in the ternary plots, we next examined specific sites with high 5hmC levels. Although the distribution of 5hmC in the ~20 million called sites varied as a function of genomic features and regulatory elements (**Fig. 5d**), the subset of CGs showing relatively high 5hmC modification levels (ACE-seq signal > 60%) was more enriched in transcriptionally active genic regions (especially introns) and less enriched in intergenic regions (**Supplementary Fig. 10**). Consistent with this observation, high-level 5hmC sites were overrepresented in actively transcribed exons (observed/random (o/r) = 2.47) and active introns (o/r = 2.57) compared with other genomic regions (**Fig. 5e**).

Finally, we rank-ordered the expression level of annotated genes and compared major histone modifications with the true 5hmC and 5mC profiles at each gene locus in purified mouse cortical excitatory neurons (**Fig. 5f**). Notably, actively transcribed genes, marked by higher H3K4me3 and lower H3K27me3 levels at promoters, showed higher intragenic 5hmC/5mC ratios than their inactive counterparts. These observations suggest that TET-mediated oxidation of 5mC in genic regions may be positively correlated with gene activity. Overall, ACE-seq analysis provides a framework for deconvolution of BS-seq data with high confidence to parse the roles of 5hmC and 5mC from an otherwise masked signal.

DISCUSSION

Recent work has demonstrated the potential for exploiting specific DNA-modifying enzymes to localize genomic features, as exemplified by the use of Tn5 transposase to localize open chromatin in ATAC-seq and the monitoring of DNA polymerase kinetics in SMRT sequencing[36,37]. In the study of DNA cytosine modifications, although great gains have been made using bisulfite-based approaches, the excessive degradation of gDNA samples has remained the most substantial limitation of chemical deamination. We took advantage of the substrate selectivity of DNA deaminases to devise an enzymatic method for base-resolution localization of 5hmC. The base-resolution maps of 5hmC generated by ACE-seq in excitatory neurons correlated strongly with those from prior TAB-seq studies on whole brain cortex, giving confidence to both techniques, as the orthogonal approaches generated similar output. The major advantage of ACE-seq is that enzymatic deamination permitted the generation of base-resolution 5hmC maps with up to 1,000-fold less DNA input. The non-destructive nature of ACE-seq was confirmed by generating both long and short amplicons from gDNA with equal efficiency after exposure of substrate gDNA to our procedure. This finding suggests that ACE-seq has the potential to profile DNA from even more rare populations, such as those in early development or cell-free DNA samples, when coupled to advances that permit library construction from limited samples. In addition, it may now be possible to examine read-lengths that are inaccessible to bisulfite-based methods, such as multi-kilobase enhancer regions.

An added attribute of ACE-seq is the robustness of discrimination of 5hmC from cytosine and 5mC, which permits high-confidence profiling of 5hmC. The rate of cytosine conversion was similar to bisulfite, whereas 5mC conversion and 5hmC non-conversion exceeded those of established methods. The resulting statistical framework allowed us to demonstrate confidently that 5hmC is a rare modification in CH contexts, a question of importance given the purported differences in 5hmC interactions with DNA binding partners such as MeCP2 in CG versus CH contexts[38]. Merging the base-resolution profiles from BS-seq and ACE-seq in purified excitatory neurons revealed the heterogeneity in 5hmC/5mC signals as a function of different genomic features and regulatory elements, where the functional effect of 5hmC enrichment and altered 5hmC/5mC ratios can now be better parsed. For example, in excitatory neurons, enhancers offer an example of a genomic feature with marked heterogeneity, but overall high 5hmC/5mC ratios, whereas ICRs offer a notable contrast, as they are mainly comprised of 5hmC-depeleted regions with low 5hmC/5mC ratios. At present, bisulfite treatment remains useful for distinguishing cytosine and 5mC; however, we envision that, building on this precedent, other schemes integrating APOBEC-mediated deamination could be optimized to discriminate between cytosine and 5mC (**Supplementary Fig. 11**). Overall, ACE-seq expands the repertoire of biotechnological approaches in which exploiting nature's toolbox of DNA-modifying enzymes can be used to great effect for characterizing and manipulating genomic DNA.

METHODS

Methods, including statements of data availability and any associated accession codes and references, are available in the online version of the paper.

Note: Any Supplementary Information and Source Data files are available in the online version of the paper.

ACKNOWLEDGMENTS

We are grateful to Z. Zhou, M. Fasolino, A. Bryson and J.M. SanMiguel for discussion and reagents. This work was supported by the US National Institutes of Health through R21-HG009545 (to R.M.K.) and by the Penn Epigenetics Institute. Additional support included R00-HG007982 (to H.W.), DP2-HL142044 (to H.W.) and R01-GM118501 (to R.M.K.). E.K.S. and J.E.D. are NSF Graduate Research Fellows. J.E.D. and E.B.F. were supported by NIH training grant T32-GM07229, and M.Y.L. by F30-CA196097.

AUTHOR CONTRIBUTIONS

E.K.S., C.S.N., R.M.K. and H.W. conceived of and developed the ACE-seq approach. E.K.S. conducted all of the experiments, with assistance from J.E.D., M.Y.L., E.B.F., P.H. and Y.H. F.D.B. contributed to phage experiment design.

H.W. performed computational analysis. E.K.S., H.W. and R.M.K. analyzed the results and wrote the manuscript, with contributions from all of the authors.

COMPETING INTERESTS

Aspects of the ACE-seq protocol have been non-exclusively licensed.

1. Schübeler, D. Function and information content of DNA methylation. *Nature* **517**, 321–326 (2015).
2. Tahiliani, M. *et al.* Conversion of 5-methylcytosine to 5-hydroxymethylcytosine in mammalian DNA by MLL partner TET1. *Science* **324**, 930–935 (2009).
3. Ito, S. *et al.* Tet proteins can convert 5-methylcytosine to 5-formylcytosine and 5-carboxylcytosine. *Science* **333**, 1300–1303 (2011).
4. He, Y.F. *et al.* Tet-mediated formation of 5-carboxylcytosine and its excision by TDG in mammalian DNA. *Science* **333**, 1303–1307 (2011).
5. Pfaffeneder, T. *et al.* The discovery of 5-formylcytosine in embryonic stem cell DNA. *Angew. Chem. Int. Edn Engl.* **50**, 7008–7012 (2011).
6. Kohli, R.M. & Zhang, Y. TET enzymes, TDG and the dynamics of DNA demethylation. *Nature* **502**, 472–479 (2013).
7. Wagner, M. *et al.* Age-dependent levels of 5-methyl-, 5-hydroxymethyl-, and 5-formylcytosine in human and mouse brain tissues. *Angew. Chem. Int. Edn Engl.* **54**, 12511–12514 (2015).
8. Bachman, M. *et al.* 5-Formylcytosine can be a stable DNA modification in mammals. *Nat. Chem. Biol.* **11**, 555–557 (2015).
9. Wu, H. & Zhang, Y. Charting oxidized methylcytosines at base resolution. *Nat. Struct. Mol. Biol.* **22**, 656–661 (2015).
10. Huang, Y. *et al.* The behavior of 5-hydroxymethylcytosine in bisulfite sequencing. *PLoS One* **5**, e8888 (2010).
11. Yu, M. *et al.* Base-resolution analysis of 5-hydroxymethylcytosine in the mammalian genome. *Cell* **149**, 1368–1380 (2012).
12. Booth, M.J. *et al.* Quantitative sequencing of 5-methylcytosine and 5-hydroxymethylcytosine at single-base resolution. *Science* **336**, 934–937 (2012).
13. Tanaka, K. & Okamoto, A. Degradation of DNA by bisulfite treatment. *Bioorg. Med. Chem. Lett.* **17**, 1912–1915 (2007).
14. Luo, C. *et al.* Single-cell methylomes identify neuronal subtypes and regulatory elements in mammalian cortex. *Science* **357**, 600–604 (2017).
15. Clark, S.J., Lee, H.J., Smallwood, S.A., Kelsey, G. & Reik, W. Single-cell epigenomics: powerful new methods for understanding gene regulation and cell identity. *Genome Biol.* **17**, 72 (2016).
16. Booth, M.J., Raiber, E.A. & Balasubramanian, S. Chemical methods for decoding cytosine modifications in DNA. *Chem. Rev.* **115**, 2240–2254 (2015).
17. Zahid, O.K., Zhao, B.S., He, C. & Hall, A.R. Quantifying mammalian genomic DNA hydroxymethylcytosine content using solid-state nanopores. *Sci. Rep.* **6**, 29565 (2016).
18. Chavez, L. *et al.* Simultaneous sequencing of oxidized methylcytosines produced by TET/JBP dioxygenases in Coprinopsis cinerea. *Proc. Natl. Acad. Sci. USA* **111**, E5149–E5158 (2014).
19. Sun, Z. *et al.* High-resolution enzymatic mapping of genomic 5-hydroxymethylcytosine in mouse embryonic stem cells. *Cell Rep.* **3**, 567–576 (2013).
20. Sun, Z. *et al.* A sensitive approach to map genome-wide 5-hydroxymethylcytosine and 5-formylcytosine at single-base resolution. *Mol. Cell* **57**, 750–761 (2015).
21. Mellén, M., Ayata, P. & Heintz, N. 5-hydroxymethylcytosine accumulation in postmitotic neurons results in functional demethylation of expressed genes. *Proc. Natl. Acad. Sci. USA* **114**, E7812–E7821 (2017).
22. Gross, J.A. *et al.* Characterizing 5-hydroxymethylcytosine in human prefrontal cortex at single base resolution. *BMC Genomics* **16**, 672 (2015).
23. Lister, R. *et al.* Global epigenomic reconfiguration during mammalian brain development. *Science* **341**, 1237905 (2013).
24. Siriwardena, S.U., Chen, K. & Bhagwat, A.S. Functions and malfunctions of mammalian DNA-cytosine deaminases. *Chem. Rev.* **116**, 12688–12710 (2016).
25. Nabel, C.S. *et al.* AID/APOBEC deaminases disfavor modified cytosines implicated in DNA demethylation. *Nat. Chem. Biol.* **8**, 751–758 (2012).
26. Wijesinghe, P. & Bhagwat, A.S. Efficient deamination of 5-methylcytosines in DNA by human APOBEC3A, but not by AID or APOBEC3G. *Nucleic Acids Res.* **40**, 9206–9217 (2012).
27. Carpenter, M.A. *et al.* Methylcytosine and normal cytosine deamination by the foreign DNA restriction enzyme APOBEC3A. *J. Biol. Chem.* **287**, 34801–34808 (2012).
28. Schutsky, E.K., Nabel, C.S., Davis, A.K.F., DeNizio, J.E. & Kohli, R.M. APOBEC3A efficiently deaminates methylated, but not TET-oxidized, cytosine bases in DNA. *Nucleic Acids Res.* **45**, 7655–7665 (2017).
29. Bryson, A.L. *et al.* Covalent modification of bacteriophage T4 DNA inhibits CRISPR-Cas9. *MBio* **6**, e00648–e15 (2015).
30. Mo, A. *et al.* Epigenomic signatures of neuronal diversity in the mammalian brain. *Neuron* **86**, 1369–1384 (2015).
31. Warnecke, P.M. *et al.* Identification and resolution of artifacts in bisulfite sequencing. *Methods* **27**, 101–107 (2002).
32. Johnson, B.S. *et al.* Biotin tagging of MeCP2 in mice reveals contextual insights into the Rett syndrome transcriptome. *Nat. Med.* **23**, 1203–1214 (2017).
33. Hu, L. *et al.* Crystal structure of TET2-DNA complex: insight into TET-mediated 5mC oxidation. *Cell* **155**, 1545–1555 (2013).
34. Xie, W. *et al.* Base-resolution analyses of sequence and parent-of-origin dependent DNA methylation in the mouse genome. *Cell* **148**, 816–831 (2012).
35. Yin, Y. *et al.* Impact of cytosine methylation on DNA binding specificities of human transcription factors. *Science* **356**, eaaj2239 (2017).
36. Buenrostro, J.D., Giresi, P.G., Zaba, L.C., Chang, H.Y. & Greenleaf, W.J. Transposition of native chromatin for fast and sensitive epigenomic profiling of open chromatin, DNA-binding proteins and nucleosome position. *Nat. Methods* **10**, 1213–1218 (2013).
37. Eid, J. *et al.* Real-time DNA sequencing from single polymerase molecules. *Science* **323**, 133–138 (2009).
38. Gabel, H.W. *et al.* Disruption of DNA-methylation-dependent long gene repression in Rett syndrome. *Nature* **522**, 89–93 (2015).

ONLINE METHODS

Development of ACE-seq protocol using T4 phage genomic DNA. WT human APOBEC3A was cloned, expressed and purified as described previously[28], with additional details provided in **Supplementary Protocol 1**. For validation of ACE-seq on T4 phage variants, genomic DNA was extracted from WT T4 phage (T4-ghmC, reference genome NCBI GenBank KJ477684.1), 147 T4 phage (T4-hmC; NCBI GenBank KJ477685.1) and GT7 T4 phage (T4-C, NCBI GenBank KJ477686.1). Modification content was validated by restriction digest and single-molecule PacBio sequencing in prior work[29], as well as by LC-MS/MS in this study, using the parameters described below (**Supplementary Fig. 2a**). 1 ng of each gDNA sample in a volume of 5 µL was heated to 95 °C for 5 min in the presence or absence of 10% DMSO and immediately snap-frozen in a dry ice/acetone bath. Then, the samples were incubated with 5 µM A3A at 37 °C for 1 h, 25 °C for 1 h, or under ramping temperature conditions from 4 to 50 °C over 2 h (**Supplementary Fig. 3**), in final buffer conditions containing 20 mM Tris, pH 6.5 and 0.1% Tween-20. A 550-bp amplicon was then PCR amplified from each reaction using Taq polymerase and primers optimized for bisulfite-converted DNA. The resulting amplicons were gel purified (Zymo Gel Extraction Kit) and TA cloned (Invitrogen TA Cloning Kit for Sequencing). Individual clones were Sanger sequenced and analyzed for C to T transitions across the amplicon.

Preparation of phage controls. For whole-genome analyses of phage DNA samples, phage DNA was enzymatically methylated at CG sites by incubating with the CG methyltransferase M.SssI (ThermoFisher Scientific) and *S*-adenosylmethionine (SAM) at 37 °C. After 4 h, additional enzyme and SAM were added to the reaction, and incubation continued for another 4 h before purification (Zymo Genomic DNA Clean and Concentrator). Restriction digest with HpaII (NEB) as well as LC-MS/MS analysis (below) were used to assess methylation status (**Supplementary Fig. 2b**). The methylated phage DNA was combined in equal amounts with T4-hmC phage and together they were sheared to ~300 bp using a Covaris sonicator (20% duty factor, 200 cycles per burst, 150 s) and the fragment sizes were characterized using a Bioanalyzer.

Preparation of mammalian gDNA. J1 WT ESCs were cultured in feeder-free gelatin-coated plates in Dulbecco's Modified Eagle Medium (DMEM) (GIBCO) supplemented with 15% FBS (GIBCO), 2 mM L-glutamine (GIBCO), 0.1 mM 2-mercaptoethanol (Sigma), nonessential amino acids (GIBCO), and 1,000 units/mL LIF (Millipore, ESG1107), 3 µM of CHIR99021 (Stemgent) and 1 µM of PD032591 (Stemgent). The culture was passaged every 2–3 d using 0.05% Trypsin (GIBCO). TET triple knockout (TKO) mESCs were generated as previously described[39,40]. Cortical excitatory neuronal nuclei (*NeuroD6/NEX*+) were purified from mouse brain as previously described[32].

Optimized ACE-seq protocol for whole-genome sequencing. Phage-only samples were analyzed with 1 ng each of pooled methylated phage and T4-hmC gDNA (**Fig. 2d**). For all mammalian DNA samples, a total of 2 or 20 ng of sheared gDNA (~300 bp) was analyzed, containing the sheared methylated phage and T4-hmC spike-in controls (0.5% total by mass). In a total volume of 5 µl, 1–20 ng of the gDNA mixture was glucosylated using UDP-glucose and T4 β-glucosyltransferase (βGT, NEB) at 37 °C for 1 h. 1 µl of DMSO was added and the sample was denatured at 95 °C for 5 min and snap cooled by transfer to a PCR tube rack pre-incubated at −80 °C. Before thawing, reaction buffer was overlaid to a final concentration of 20 mM MES pH 6.0 + 0.1% Tween, and A3A was added to a final concentration of 5 µM in a total volume of 10 µl. The deamination reactions were incubated under linear ramping temperature conditions from 4–50 °C over 2 h. After deamination, the samples were prepared for Illumina sequencing using the Accel Methyl-NGS kit (Swift Biosciences). The resulting ACE-seq libraries were sequenced at 1.9 pM with single-end mode on a NextSeq 500 sequencer (Illumina) using the NextSeq 500/500 High Output kit v2 (150 cycles). A more detailed step-by-step protocol for ACE-seq, with added rationale and discussion, is provided as **Supplementary Protocol 2**.

LC-MS/MS analysis of phage genome controls. To determine the purity of the phage gDNA stocks and the efficiency of both the glucosylation and deamination steps in ACE-seq, samples were analyzed by LC-MS/MS. For the phage validation experiments (**Supplementary Fig. 2**), untreated phage gDNA was used. For the experiments analyzing ACE-seq efficiency (**Fig. 2e** and **Supplementary Fig. 4**), 15 ng of the phage-only samples (pooled samples with 7.5 ng each of methylated phage and T4-hmC gDNA) were treated using the ACE-seq protocol, excluding either βGT or A3A, or excluding both βGT and A3A. The gDNA samples were degraded with 1 U DNA Degradase Plus (Zymo) in 10–15 µl at 37 °C for 4 h. The nucleoside mixture was diluted tenfold into 0.1% formic acid, and 2 ng was injected onto an Agilent 1200 Series HPLC with a 5 µm, 2.1 × 250 mm Supelcosil LC-18-S analytical column (Sigma) equilibrated to 45 °C in Buffer A (0.1% formic acid). The nucleosides were separated using a gradient of 0–10% Buffer B (0.1% formic acid, 30% (v/v) acetonitrile) over 8 min at a flow rate of 0.5 ml/min. Tandem MS/MS was performed in positive ion mode ESI on an Agilent 6460 triple-quadrupole mass spectrometer, with gas temperature of 225 °C, gas flow of 12 L/min, nebulizer at 35 psi, sheath gas temperature of 300 °C, sheath gas flow of 11 L/min, capillary voltage of 3,500 V, fragmentor voltage of 70 V, and delta EMV of +1,000 V. Collision energies were optimized to 10 V for C, U, T and 5mC, and 25 V for 5hmC and 5ghmC. MRM mass transitions were C C 228.1 -›112.1 *m/z*, U 229.1 -›113.0, T 243.1 -›127.1, 5mC 242.1 -›126.1; 5hmC 258.1 -›124.1; and 5ghmC 420.2 -›124.1. Standard curves were generated from standard nucleosides (Berry & Associates) ranging from 10 pmol to 2.4 fmol for all nucleosides, with the exception of 5ghmC for which a standard is not readily available. When possible, the sample peak areas were fit to the standard curve to determine amounts of each modified cytosine in the gDNA sample. To account for slight variations in the amount of material loaded, the calculated concentrations (or raw peak areas for 5ghmC) were normalized to that of G, allowing us to make comparisons between samples.

Input DNA comparison of ACE-seq and BS-seq by fixed-cycle PCR or by qPCR. Half log dilutions of WT ESC DNA (from 1 µg to 1 ng) were subjected to standard bisulfite treatment (Qiagen EpiTect Bisulfite Kit) or to ACE-seq, following the optimized protocol with two exceptions: first, the samples contained WT ESC gDNA that was not sheared and did not contain phage spike-in DNA, and, second, the samples were not subjected to library preparation after the deamination reactions. For fixed-cycle PCR analysis, after the deamination procedure, 0.5 or 1 µl of each reaction (normalized to contain an equivalent amount of starting template DNA) was used to seed two different PCR reactions: Either short (200 bp) and long (1 kb) fragments were amplified from the *Tbx5* genomic region using primers designed with a bisulfite-optimized algorithm (MethPrimer) and KAPA Hifi HotStart Uracil+ ReadyMix (KAPA Biosystems). After 35 cycles of PCR, resulting products were visualized on a 1.5% agarose gel stained with SybrSafe (**Fig. 3a,b**). For qPCR analysis (**Fig. 3c** and **Supplementary Fig. 5**), 0.5 µl of the treated samples were combined with 500 nM each of the forward and reverse primers, to amplify either the 200-bp amplicon or the 1-kb amplicon, and amplified using the KAPA SYBR Fast Rox low qPCR Mastermix kit (KAPA Biosystems). For the 200-bp amplicon, a two-step PCR protocol was used in which the samples were initially denatured at 95 °C for 3 min, and then cycled between 95 °C (15 s) and 63 °C (20 s) for a total of 35 cycles. For the 1-kb amplicon, a two-step PCR protocol was used in which the samples were initially denatured at 95 °C for 5 min, and then cycled between 95 °C (30 s) and 66 °C (90 s) for a total of 41 cycles. Calculated C_T values for each sample were normalized to the C_T values calculated in the no template controls (resulting from primer dimer signal), and one cycle was subtracted from ACE-seq measurements to account for differences in gDNA input associated with sample dilution. Resulting qPCR products were run on 1% agarose gels and stained with SybrSafe to confirm specific amplification (**Supplementary Fig. 5**); notably, the background signals from the 'no template' and bisulfite-treated samples for the 1-kb amplicon were exclusively from primer dimer amplification and were not specific to the desired 1-kb product.

Data processing for whole-genome ACE-seq. Sequencing reads were processed as previously reported[41]. Briefly, raw reads were trimmed for low-quality bases and adaptor sequences using Trim Galore (https://www.bioinformatics.babraham.ac.uk/projects/trim_galore/), and the data quality was examined with FastQC (http://www.bioinformatics.babraham.ac.uk/projects/fastqc/). The trimmed reads were mapped against the reference genomes with Bismark

(v0.14.3)[42]. PCR duplicates were removed using the Picard program (http://picard.sourceforge.net/). To eliminate reads from strands not deaminated by A3A, reads with three or more consecutive non-converted cytosines in the CH context were removed. Raw signals were calculated as % of C/(C+T) at each site. Statistics of all genome-scale sequencing libraries are summarized in **Supplementary Table 1**.

Statistical calling of 5hmC and assessing FDR of whole-genome ACE-seq. For each cytosine in CG dinucleotides, we counted the number of 'C' bases from ACE-seq reads as 5hmC (denoted N_C) and the number of 'T' bases as methylated or unmodified cytosines (denoted N_T). For statistical calling, we used the binomial distribution (N as the sequencing coverage ($N_T + N_C$) and p as the error rate of A3A deamination (0.47%, averaged non-conversion rate for 5mCG in spiked-in λ phage DNA, from six independent measurements)) to assess the probability of observing N_c or greater by chance (**Fig. 4e**). To estimate empirical FDR of calling 5hmC-modified CGs, we repeated the steps above on ACE-seq signals of a negative control sample (TET TKO ESCs) in which all authentic ox-mCs are absent. The FDR for a given P value cutoff of the binomial distribution is the number of called CGs in negative controls divided by the number detected in the sample. For estimating the contribution of 5fC/5caC to ACE-seq signals, we used the global modification levels of 5hmC (~5,000 p.p.m.), 5fC (~10 ppm) and 5caC (undetectable) measured by the quantitative mass spectrometry analysis of adult mouse brain[8,43]. This suggests that the false positive signal derived from 5fC/5caC is at most 0.2%. For calling 5hmC in CH contexts (CHH or CHG), a sequence-context-matched error rate of A3A deamination (0.10%) was used to calculate empirical FDR (**Fig. 4e**). We restricted our statistical analysis to CG, CHH or CHG sites covered by at least five reads per strand.

Pairwise comparisons with ACE-seq. For pairwise comparison between ACE-seq and TAB-seq in neurons, the raw 5hmCG signals were calculated within tiled 10-kb genomic bins (**Fig. 4b**). For pairwise comparison between ACE-seq performed with 2 ng versus 20 ng of gDNA, the raw 5hmCG signals were similarly calculated within tiled 10-kb genomic bins (**Supplementary Fig. 6**). Pearson correlation coefficients were calculated using R function *cor*.

Calculating the true level of 5mCGs by combining ACE-seq with BS-seq. For each CG site, the levels of 5mC and 5hmC were estimated using the MLML tool[44]. This approach arrives at maximum likelihood estimates for the 5mC and 5hmC levels by combining data from ACE-seq and BS-seq (see below). Only CG sites with 0 conflicts were considered for further analysis. From the MLML output, the level of unmodified CG was estimated by [100% – (abundance of 5hmC + 5mC)]. The results were further filtered, such that CG, 5mCG, and 5hmCG levels were non-negative. For generating ternary plots (**Fig. 5** and **Supplementary Fig. 8**), levels of CG, 5mCG, and 5hmCG (as percentage of the sum of [CG + 5mCG + 5hmCG]) were calculated within tiled 1-kb genomic bins across the genome.

Quantifying enrichment of called 5hmCGs at genomic elements. To calculate the enrichment of statistically confident 5hmCGs at a set of genomic elements (**Fig. 5e** and **Supplementary Fig. 10**), we counted the number of overlapping 5hmCGs and divided by the average of ten random samplings of called 5hmCGs. The sampling involves the shuffling of genomic elements within the same chromosome. We then normalized this enrichment value by the genomic span of the corresponding set of genomic elements.

Genome browser visualization. We used Integrative Genomics Viewer (IGV, v2.3.91)[45] to visualize ACE-seq signals using mm9 (**Fig. 4a**) or mm10 (**Fig. 5a** and **Supplementary Fig. 7**) Refseq transcript annotation as reference. For ACE-seq, TAB-seq, and BS-seq data sets, modified cytosines are indicated by upward (plus strand) and downward (minus strand) ticks, with the height of each tick representing the fraction of modification at the site ranging from 0 to 1. For RNA-seq, ATAC-seq and ChIP-seq data, read density was normalized to 10 million reads.

Analysis of 5hmCGs and 5mCGs within imprinted regions. For **Figure 5** and **Supplementary Figure 9**, we analyzed 5hmCG and 5mCG modification levels within 30 of 32 well-established parental origin-dependent allele specific regions[34] (see **Supplementary Table 2**). The *Dlk1-Gtl2* IG and *Igf2r* loci were excluded as a result of aberrantly high 5mC levels and low coverage, respectively. The mean 5hmC levels were calculated for each imprinted region using the base-resolution ACE-seq data set of cortical excitatory neurons generated in this study. The true 5mC levels were then estimated via subtraction of the ACE-seq signal from previously-published bisulfite-sequencing data in mouse brain[34]. For imprinted regions ($n = 22$) with less than 10% mean 5hmC levels, we denote them as 5hmC-low imprinted regions in **Supplementary Figure 9**. Those ($n = 8$) with >10% mean 5hmC levels were denoted as 5hmC-high imprinted regions.

Statistical analysis. No statistical methods were used to predetermine sample size for any experiments. All group results are expressed as mean ± s.d. unless otherwise stated. Specific P values used for calling modified cytosine bases are explicitly stated in the text and figure legends. Each figure legend explicitly states the number of independent experiments. Statistical analyses for graphs were performed in GraphPad PRISM 7.

Published data sets. For **Figure 4a–c**, we used the following published data sets: 5hmC DIP-seq in mouse ESCs[46], TAB-seq in mouse ESCs[11], TAB-seq in mouse frontal cortex[23]. For **Figure 5** and **Supplementary Figure 7**, we used whole-genome BS-seq, RNA-seq, ATAC-seq and ChIP-seq for H3K4me1, H3K4me3, H3K27me3 and H3K27ac in purified mouse cortical excitatory neurons[30].

Reporting Summary. Further information on research design is available in the Nature Research Reporting Summary linked to this article.

Data availability. All sequencing data associated with this study have been deposited to Gene Expression Omnibus (GEO) under the accession code GSE116016.

39. Wang, H. *et al.* One-step generation of mice carrying mutations in multiple genes by CRISPR/Cas-mediated genome engineering. *Cell* **153**, 910–918 (2013).
40. Lu, F., Liu, Y., Jiang, L., Yamaguchi, S. & Zhang, Y. Role of Tet proteins in enhancer activity and telomere elongation. *Genes Dev.* **28**, 2103–2119 (2014).
41. Wu, H., Wu, X., Shen, L. & Zhang, Y. Single-base resolution analysis of active DNA demethylation using methylase-assisted bisulfite sequencing. *Nat. Biotechnol.* **32**, 1231–1240 (2014).
42. Krueger, F. & Andrews, S.R. Bismark: a flexible aligner and methylation caller for Bisulfite-Seq applications. *Bioinformatics* **27**, 1571–1572 (2011).
43. Bachman, M. *et al.* 5-Hydroxymethylcytosine is a predominantly stable DNA modification. *Nat. Chem.* **6**, 1049–1055 (2014).
44. Qu, J., Zhou, M., Song, Q., Hong, E.E. & Smith, A.D. MLML: consistent simultaneous estimates of DNA methylation and hydroxymethylation. *Bioinformatics* **29**, 2645–2646 (2013).
45. Thorvaldsdóttir, H., Robinson, J.T. & Mesirov, J.P. Integrative Genomics Viewer (IGV): high-performance genomics data visualization and exploration. *Brief. Bioinform.* **14**, 178–192 (2013).
46. Shen, L. *et al.* Genome-wide analysis reveals TET- and TDG-dependent 5-methylcytosine oxidation dynamics. *Cell* **153**, 692–706 (2013).

doi:10.1038/nbt.4204

nature research

Corresponding author(s): Rahul M. Kohli; Hao Wu

Life Sciences Reporting Summary

Nature Research wishes to improve the reproducibility of the work that we publish. This form is intended for publication with all accepted life science papers and provides structure for consistency and transparency in reporting. Every life science submission will use this form; some list items might not apply to an individual manuscript, but all fields must be completed for clarity.

For further information on the points included in this form, see Reporting Life Sciences Research. For further information on Nature Research policies, including our data availability policy, see Authors & Referees and the Editorial Policy Checklist.

Please do not complete any field with "not applicable" or n/a. Refer to the help text for what text to use if an item is not relevant to your study.
For final submission: please carefully check your responses for accuracy; you will not be able to make changes later.

Experimental design

1. Sample size

Describe how sample size was determined.

No sample size pre-calculations were performed.

2. Data exclusions

Describe any data exclusions.

From next generation sequencing data, PCR duplicates were removed and the sequences further filtered to remove strands with three consecutive non-converted CpHs. The complete statistics for removal of filtered data are provided in SI Table 1. In ICR analysis, the Dlk1-Gtl2 IG and Igf2r loci were excluded due to aberrantly high 5mC levels and low coverage, respectively

3. Replication

Describe the measures taken to verify the reproducibility of the experimental findings.

The data was reliably reproduced. Comparison of different gDNA inputs is shown in SI Fig 6.

4. Randomization

Describe how samples/organisms/participants were allocated into experimental groups.

N/A

5. Blinding

Describe whether the investigators were blinded to group allocation during data collection and/or analysis.

N/A

Note: all in vivo studies must report how sample size was determined and whether blinding and randomization were used.

6. Statistical parameters

For all figures and tables that use statistical methods, confirm that the following items are present in relevant figure legends (or in the Methods section if additional space is needed).

n/a	Confirmed
☐	☒ The exact sample size (*n*) for each experimental group/condition, given as a discrete number and unit of measurement (animals, litters, cultures, etc.)
☐	☒ A description of how samples were collected, noting whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
☐	☒ A statement indicating how many times each experiment was replicated
☐	☒ The statistical test(s) used and whether they are one- or two-sided *Only common tests should be described solely by name; describe more complex techniques in the Methods section.*
☐	☒ A description of any assumptions or corrections, such as an adjustment for multiple comparisons
☐	☒ Test values indicating whether an effect is present *Provide confidence intervals or give results of significance tests (e.g. P values) as exact values whenever appropriate and with effect sizes noted.*
☐	☒ A clear description of statistics including central tendency (e.g. median, mean) and variation (e.g. standard deviation, interquartile range)
☐	☒ Clearly defined error bars in all relevant figure captions (with explicit mention of central tendency and variation)

See the web collection on statistics for biologists for further resources and guidance.

Nature Biotechnology: doi:10.1038/nbt.4204

Software

Policy information about availability of computer code

7. Software

Describe the software used to analyze the data in this study.

The software used is described in the Methods section in detail. The programs used include:
1. Trim Galore (v0.4.1): https://www.bioinformatics.babraham.ac.uk/projects/trim_galore/
2. FastQC (v0.11.5): https://www.bioinformatics.babraham.ac.uk/projects/fastqc/
3. Bismark (v0.14.3): https://www.bioinformatics.babraham.ac.uk/projects/bismark/
4. Picard (v2.5.0): https://broadinstitute.github.io/picard/
5. methpipe/MLML (v3.4.3): http://smithlabresearch.org/software/mlml/
6. R/ggplot2 (v2.2.1): http://ggplot2.tidyverse.org/
7. Integrative Genomics Viewer (v2.3.32): https://software.broadinstitute.org/software/igv/

For manuscripts utilizing custom algorithms or software that are central to the paper but not yet described in the published literature, software must be made available to editors and reviewers upon request. We strongly encourage code deposition in a community repository (e.g. GitHub). *Nature Methods* guidance for providing algorithms and software for publication provides further information on this topic.

Materials and reagents

Policy information about availability of materials

8. Materials availability

Indicate whether there are restrictions on availability of unique materials or if these materials are only available for distribution by a third party.

All materials are available commercially, with the exception of APOBEC3A and genomic DNA derived from cell lines or mice. The APOBEC3A expression plasmid has been deposited in Addgene for release upon publication to academic users or can be requested from authors.

9. Antibodies

Describe the antibodies used and how they were validated for use in the system under study (i.e. assay and species).

No antibodies were used.

10. Eukaryotic cell lines

a. State the source of each eukaryotic cell line used.

ESCs (J1) were originally purchased from ATCC and no additional authentication was performed. TET TKO cells were derived from two independent methods. One sample was obtained from the Zhang lab, with generation described in Lu et al, Genes Dev. (2014) 28: 2103-2119. The second TET TKO cell line was derived in house using the method from Wang et al, Cell (2013) 153:910-918. Both cells lines were confirmed as noted below.

b. Describe the method of cell line authentication used.

TET TKO ESCs were verified by Sanger sequencing at Tet1, Tet2 and Tet3. The lack of hmC was confirmed by mass spectrometry.

c. Report whether the cell lines were tested for mycoplasma contamination.

Not tested.

d. If any of the cell lines used are listed in the database of commonly misidentified cell lines maintained by ICLAC, provide a scientific rationale for their use.

None of the cell lines used were listed in ICLAC.

Animals and human research participants

Policy information about studies involving animals; when reporting animal research, follow the ARRIVE guidelines

11. Description of research animals

Provide all relevant details on animals and/or animal-derived materials used in the study.

Purified mouse neuronal gDNA was provided by Z. Zhou, as per Johnson BS et al, Nat Med (2017) 23:1203-1214. The samples were taken from 6-week old C57BL/6 male mice.

Policy information about studies involving human research participants

12. Description of human research participants

Describe the covariate-relevant population characteristics of the human research participants.

This study did not involve human participants.

Nature Biotechnology: doi:10.1038/nbt.4204