# Evaluation of CA repeat arrays as targets for MeCP2 function in the brain

A Preprint

**Kashyap Chhatbar**[*,1,2], **John Connelly**[*,1], **Shaun Webb**[1], **Skirmantas Kriaucionis**[3], **and Adrian Bird**[1,✉]

[*]These authors contributed equally to this manuscript
[1]Wellcome Centre for Cell Biology, University of Edinburgh, The Michael Swann Building, Max Born Crescent, Mayfield Road, Edinburgh EH9 3BF, UK
[2]Informatics Forum, School of Informatics, University of Edinburgh, 10 Crichton Street, Edinburgh EH8 9AB, UK
[3]Ludwig Institute for Cancer Research, University of Oxford, Oxford, UK
✉Correspondence: a.bird@ed.ac.uk

## Abstract

The DNA binding protein MeCP2 is reported to bind methylated cytosine in CG and CA motifs in genomic DNA, but it was recently proposed that arrays of tandemly repeated CA containing either methylated or hydroxymethylated cytosine are the primary targets for MeCP2 binding and function. Here we investigated the predictions of this hypothesis using a range of published datasets. We failed to detect enrichment of cytosine modification at genomic CA repeat arrays in mouse brain regions and found no evidence for preferential MeCP2 binding at CA repeats. Moreover, we did not observe a correlation between the CA repeat density near genes and their degree of transcriptional deregulation when MeCP2 was absent. Our results do not provide support for the hypothesis that CA repeats are key mediators of MeCP2 function. Instead, we found that CA repeats are subject to CAC methylation to a degree that is typical of the surrounding genome and contribute modestly to MeCP2-mediated modulation of gene expression in accordance with their content of this canonical target motif.

## Introduction

MeCP2 is a chromatin protein that is abundant in neurons and essential for brain function. Initially identified through its affinity for 5-methylcytosine in DNA, the exact nature of its DNA targets has periodically been debated and challenged. Based on in vivo and in vitro data from several laboratories, targeting of MeCP2 to DNA depends on the presence of mC in two motifs: mCG and mCA (the latter mostly in the trinucleotide mCAC) (Guo et al., 2014; Gabel et al., 2015; Lagger et al., 2017; Lewis et al., 1992; Cholewa-Waclaw et al., 2019; Connelly et al., 2020; Ayata, 2013). Independent studies suggest that the presence of mCA in neurons is essential, in part at least because of its affinity for MeCP2 (Gabel et al., 2015; Tillotson et al., 2021; Lavery et al., 2020). For example, mice expressing a modified form of MeCP2 that does not interact with mCA but can still bind mCG develop severe Rett syndrome-like phenotypes (Tillotson et al., 2021). In vitro, MeCP2 also binds to the hydroxymethylated (hm) motif hmCAC (Lagger et al., 2017), but the significance of this interaction has been unclear due to the apparent rarity of this modified trinucleotide in brain genomes (Lister et al., 2013).

This scenario is challenged by a recent proposal that arrayed tandem repeats of the dinucleotide CA are critical MeCP2 targets, exceeding in importance both mCG and isolated mCA moieties as mediators of MeCP2 function (Ibrahim et al., 2021). As $[CA]_n$ repeat blocks are relatively frequent, it is argued that their proximity to genes provides high affinity 'landing pads' through

which MeCP2-dependent gene regulation is mediated. Mechanistically, it is suggested that MeCP2 binding to occasional mCA or hmCA moieties within $[CA]_n$ repeats can seed cooperative MeCP2 binding across the entire array, including non-methylated CA motifs. Here, we investigate further the relationship between MeCP2 and CA repeats. Our findings do not offer support for the claim that cytosine modification is enriched at $[CA]_n$ arrays or that MeCP2 is preferentially bound at these repeat blocks in brain cell nuclei. Moreover, we find that the effects of MeCP2-deficiency on transcription in various brain regions do not correlate with proximity to $[CA]_n$ repeats, but instead strongly correlate with local mCAC frequency.

# Results

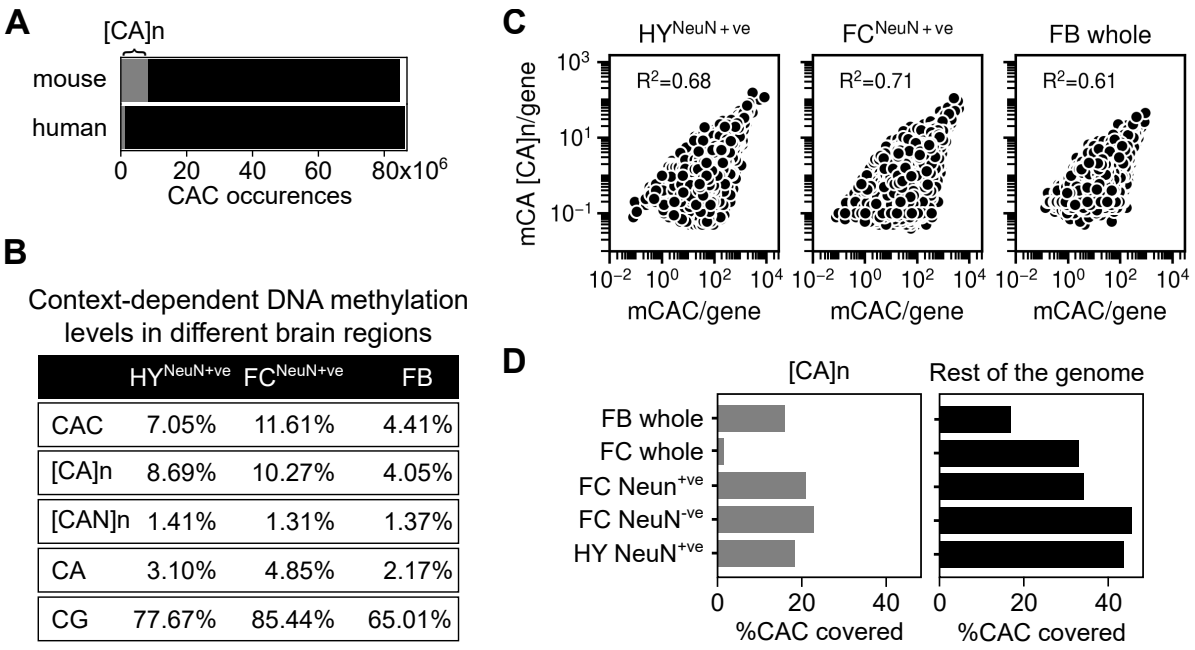
## Absence of enrichment of modified cytosine in CA repeats

The mouse genome (version mm9) contains ~320,000 $[CA]_n$ arrays (minimum 10 base pairs) of variable length with an average of 25 CA repeats each. It has been reported that CA is more frequently methylated or hydroxymethylated within $[CA]_n$ repeat arrays than elsewhere in the genome (Ibrahim et al., 2021), but this comparison did not take account of the preferential methylation of CAC trinucleotides in neurons (Lagger et al., 2017). While CAC is necessarily very abundant within $[CA]_n$ repeats, isolated CA motifs elsewhere in the genome may or may not have C in the third position. Recognising that the trinucleotide sequence CAC is the preferred target of non-CG methylation in neurons and is also a target for MeCP2 binding (Lagger et al., 2017; de Mendoza et al., 2021), we determine that approximately 8% of all CAC motifs in mouse are found within $[CA]_n$ repeats (Figure 1A). Interestingly, humans (version chm13-v1.1) possess a much lower proportion of $[CA]_n$ repeats: ~55,000 $[CA]_n$ arrays amounting to only ~1% of all CAC motifs in the genome (Figure 1A). Using published data for three mouse brain regions, we confirmed that mCA occurs at a lower frequency outside $[CA]_n$, but the frequency of mCAC for each brain region was similar within and outside the repeat arrays (Figure 1B). To determine whether the levels of cytosine modification in $[CA]_n$ arrays match the level of mCAC nearby, we plotted mCAC number per gene against mCA number exclusively within CA repeats (Figure 1C). The results showed a strong correlation, indicating that the local density of mCAC in genes is similar regardless of whether the trinuclotide is isolated or within a CA repeat array. We conclude from these findings that, while $[CA]_n$ arrays are subject to CAC methylation, they are not targeted preferentially compared with the surrounding genome but tend to adopt a level of mCAC that reflects the surrounding DNA.

This conclusion assumes that the level of cytosine modification in $[CA]_n$ arrays has not been systematically underestimated by bisulfite sequencing due to reduced coverage of repetitive sequences. To test for under-representation in published bisulfite sequence data, we compared the fraction of CAC covered in $[CA]_n$ repeats versus the rest of the genome. The results show little difference for forebrain and, moderately reduced coverage of $[CA]_n$ arrays (less than two-fold) in $NeuN^{+ve}$ frontal cortex and hypothalamus (Figure 1D). It is noteworthy that bisulfite sequence data for whole frontal cortex gave low coverage of $[CA]_n$ repeats, but this effect was largely absent in purified cortical $NeuN^{+ve}$ (neuronal) or $NeuN^{-ve}$ (mostly non-neuronal) nuclei from the same study (Lister et al., 2013) (Figure 1D). The evidence indicates that bisulfite sequencing does not intrinsically discriminate against $[CA]_n$ repeats and therefore the level of mC in these arrays has not been greatly underestimated.

## Absence of enrichment of MeCP2 binding at CA repeats in brain

We next asked whether MeCP2 is preferentially associated with CA repeat arrays (Ibrahim et al., 2021). MeCP2 ChIP of mouse brain reproducibly reveals relatively uniform genome occupancy with few prominent peaks (Gabel et al., 2015; Lagger et al., 2017; Chen et al., 2015). This has been interpreted to reflect the high frequency throughout the neuronal genome of short MeCP2 target sites, mCG and mCAC (Lagger et al., 2017). In contrast, Ibrahim et al report that MeCP2 ChIP-seq reads in cultured mouse embryo fibroblasts (MEFs) are concentrated in prominent peaks coincident with CA repeat clusters (Ibrahim et al., 2021). Given the importance of MeCP2 function and the exceptionally high abundance of hmC and mC in neurons, equivalent peaks at $[CA]_n$ might be expected in the brain. However, published ChIP data does not support this prediction. As an example, the MeCP2 binding profile (normalised to the KO or input ChIP profiles) across the same ~120 kb region of the mouse genome that was illustrated for MEFs (Ibrahim et al., 2021) failed to highlight $[CA]_n$ arrays (Figure 2A). In view of




| | HY NeuN+ve | FC NeuN+ve | FB |
|---|---|---|---|
| CAC | 7.05% | 11.61% | 4.41% |
| [CA]n | 8.69% | 10.27% | 4.05% |
| [CAN]n | 1.41% | 1.31% | 1.37% |
| CA | 3.10% | 4.85% | 2.17% |
| CG | 77.67% | 85.44% | 65.01% |

Figure 1 **Absence of enrichment of modified cytosines in CA repeat arrays. A)** Number of CAC occurrences across the mouse and human genome. The grey area corresponds to the number of CAC occurrences within CA repeat arrays. **B)** Genome wide average DNA methylation levels for different cytosine contexts: CAC; $[CA]_n$ dinucleotide repeats; [CAN]n trinucleotide repeats (where N is A, T, or G); CA; and CG. DNA methylation levels were quantified from three mouse brain regions: sorted $NeuN^{+ve}$ nuclei from hypothalamus (HY) (Lagger et al., 2017); sorted $NeuN^{+ve}$ nuclei from frontal cortex (FC) (Lister et al., 2013) and forebrain (FB) (Boxer et al., 2020). **C)** mCAC number per gene plotted against mCA number per gene exclusively in CA repeats and Pearson correlation ($R^2$) is calculated for three mouse brain regions: sorted $NeuN^{+ve}$ nuclei from hypothalamus (HY) (Lagger et al., 2017); sorted $NeuN^{+ve}$ nuclei from frontal cortex (FC) (Lister et al., 2013) and forebrain (FB) (Boxer et al., 2020). **D)** Percentage of CAC sites with adequate bisulfite sequence coverage within CA repeat loci and across the mouse genome. The coverage threshold for sorted $NeuN^{+ve}$ Hypothalamus (HY) (Lagger et al., 2017) and whole forebrain (FB) (Boxer et al., 2020) is at least 5 reads and threshold for sorted $NeuN^{+ve}$, $NeuN^{-ve}$ and whole frontal cortex (FC) (Lister et al., 2013) is at least 10 reads.

uncertainties regarding the initial MEF data (see Discussion), our findings question the evidence for preferential localisation of MeCP2 to CA repeats.

We also visualised enrichment of MeCP2 at lower resolution across a much larger (15 megabase) region of mouse chromosome 5 (Figure 2B). The results agree with previous reports of a fluctuating distribution of MeCP2 across the genome that broadly tracks the density of mCAC (Lagger et al., 2017). Although $[CA]_n$ repeat arrays are not apparent as prominent sites of MeCP2 binding, their high density at this resolution makes it difficult to discern by inspection alone whether they are preferred. In addition, the global distribution of bound MeCP2 across the neuronal genome limits the value of traditional peak analysis methods to define the binding pattern (Lagger et al., 2017). To reveal the relationship of ChIP signal to CA repeats versus other genomic regions, we plotted the $\log_2$-fold change in MeCP2 binding across 1 kb windows (normalised to KO or input) versus mCAC frequency for three DNA sequence categories: i) $[CA]_n$; ii) repeated sequences excluding $[CA]_n$; and iii) the rest of genome excluding $[CA]_n$. We drew upon published datasets derived from brain regions for which matching ChIP and bisulfite data were available and again normalised to KO ChIP signal or input (Lagger et al., 2017; Lister et al., 2013; Chen et al., 2015; Boxer et al., 2020). The results showed that for hypothalamus, MeCP2 occupancy clearly rose as mCAC frequency increased (Figure 2C). If $[CA]_n$ arrays were preferential targets for MeCP2 binding, we would expect them to show elevated ChIP enrichment compared with the other genome categories, but the three DNA sequence types showed closely similar profiles in each brain region. The same conclusion can be drawn from data from forebrain, which is represented by two

high-coverage datasets using independent anti-MeCP2 antibodies (Figure 2D). A fourth dataset derived from mouse frontal cortex (Gabel et al., 2015; Lister et al., 2013) also showed a trio of near-identical ChIP plots, but in this case the relationship to mCAC frequency was not apparent (Figure 2E). Referring to Figure 2B, we note that whereas forebrain and hypothalamus ChIP-seq plots across this extended region of the mouse genome were closely similar, as expected given their similar DNA methylation profiles, for unknown reasons the data for frontal cortex produced a very different profile when plotted in the same manner. Regardless of differences in this respect, all three datasets failed to reveal evidence of enhanced binding of MeCP2 at $[CA]_n$, as the intensity of ChIP signal at $[CA]_n$ was approximately equivalent to that in other parts of the genome.

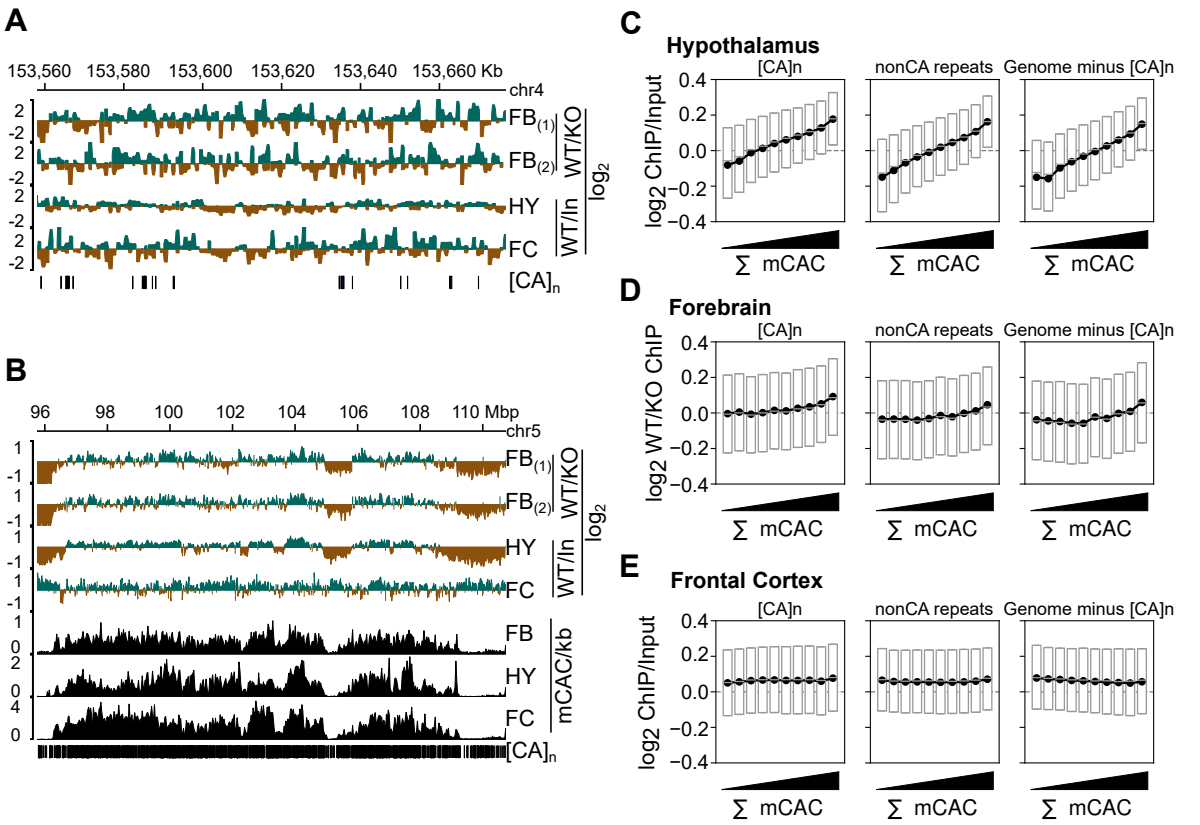



Figure 2 **Absence of enhanced MeCP2 binding at CA repeat arrays in brain.** A) Genome browser screenshots of mouse chr4:153558029-153676791 (version mm9) showing *Mecp2* WT ChIP signal normalized to *Mecp2* KO ChIP signal for forebrain (FB) using two distinct antibodies (Boxer et al., 2020), *Mecp2* WT ChIP signal normalized to input chromatin for hypothalamus (HY) (Chen et al., 2015) and frontal cortex (FC) (Gabel et al., 2015). Vertical strokes (bottom row) show location of CA repeat arrays. **B)** As described for panel A but using coordinates at chr5:95751179-110732516. Additionally, DNA methylation tracks showing mCAC/kb for sorted $NeuN^{+ve}$ nuclei from forebrain (FB) (Lister et al., 2013), hypothalamus (HY) (Lagger et al., 2017) and sorted $NeuN^{+ve}$ nuclei from frontal cortex (FC) (Lister et al., 2013). **C)** MeCP2 ChIP signal normalized to input chromatin in mouse hypothalamus (Chen et al., 2015) plotted against bins of increasing levels of DNA methylation at CAC (Lagger et al., 2017) (shown as truncated box plots) within three different genomic sequence categories. Panels from left to right represent CA repeat loci, repeat loci other than CA repeats as identified by RepeatMasker, and 500,000 randomly chosen 1kb genomic windows which do not overlap with CA repeats. **D)** As described for panel C but using data for mouse forebrain (Boxer et al., 2020). **E)** As described for panel C but using data for mouse frontal cortex (Gabel et al., 2015) and (Lister et al., 2013).

Based on electrophoretic mobility shift assays, it was further proposed that cooperative binding across $[CA]_n$ arrays is facilitated by the affinity of MeCP2 for non-methylated $[CA]_n$ (Ibrahim et al., 2021). This recalls early reports (Weitzel et al., 1997) that certain truncated variants of MeCP2 can bind to CA/TG-rich probes in vitro, although recent evidence failed to validate this mode of binding with full-length MeCP2 either in vitro or in vivo (Connelly et al., 2020). Ibrahim et al. reported MeCP2

binding to longer non-methylated CA repeat tracts – specifically $[CA]_7$. Using a pulldown assay for native MeCP2 binding in mouse brain extracts, however, we failed to detect enhanced MeCP2 binding to $[CA]_7$ compared with control DNA probes that lacked CA repeats or in which CA was part of a CAGA repeat array (Figure 3A & B). Introduction of one mC residue into the $[CA]_7$ tract significantly increased MeCP2 binding. Moreover, binding displayed a trend towards further enhancement when three more cytosines in the array were methylated. This result is compatible with a linear rather than a cooperative relationship between the amount of mCAC and MeCP2 binding, although this relationship is not quantitatively proven using our data.

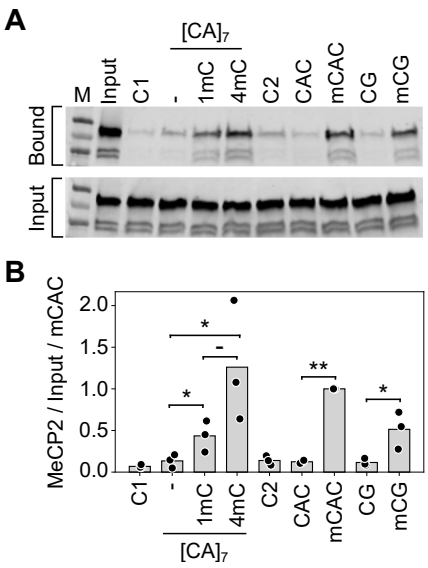



Figure 3 **MeCP2 binding to $[CA]_7$ is dependent on cytosine methylation. A)** Example of a pulldown assay for MeCP2 binding using biotin tagged double stranded DNA oligonucleotides incubated with mouse brain nuclear extracts (see Materials and Methods: Pulldown assay for MeCP2 binding to DNA). Unrelated probes C1 and C2 contained no mC. Probe CAC contained 10 non-methylated CACs on one strand, all methylated in mCAC. Alternative versions of the $[CA]_7$ probe contained 0, 1 and 4 mCAC motifs. **B)** Quantification of the triplicate data exemplified in A. Significance was estimated using a paired T-test (* pval < 0.05, ** pval < 0.01).

## Minor effect of CA repeats on MeCP2-mediated gene regulation

We next tested the relationship between gene expression changes in the MeCP2-deficient brain and the frequency of $[CA]_n$ repeat clusters within gene bodies, drawing on data from independent studies of mouse hypothalamus, forebrain and cortex (Chen et al., 2015; Boxer et al., 2020; Gabel et al., 2015). To investigate the effect of enriched $[CA]_n$ repeat clusters on transcription, we plotted the percentage of all transcription unit nucleotides that belong to $[CA]_n$ arrays against the fold change in gene expression when MeCP2 is absent. This resembles a previous analysis (Ibrahim et al., 2021) but differs by taking into account the direction of transcriptional change and testing multiple brain datasets. The results showed no obvious correlation between the differing levels of $[CA]_n$ in the gene body and changes in gene expression in these brain regions (Figure 4, left panels). In contrast, we found that the number of mCAC motifs per gene, either including or excluding $[CA]_n$ repeat clusters, correlates positively with the average magnitude of gene up-regulation in the mutant brain (Figure 4, middle panels), supporting the notion that MeCP2 binding to this methylated motif restrains gene expression and confirming previous findings (Lagger et al., 2017; Kinde et al., 2016). This strong relationship, which is unaffected by inclusion or exclusion of $[CA]_n$ arrays, is not expected if CA repeats were the primary drivers of MeCP2-mediated gene regulation. Since CA repeats are subject to DNA methylation at the same level as dispersed CAC motifs (see Figure 1), we expected that the presence mCAC within $[CA]_n$ tracts would correlate with gene expression. This was confirmed when the number of mCA motifs in $[CA]_n$ repeat blocks was plotted against the fold change in transcription between *Mecp2* KO and wildtype brain regions (Figure 4, right panels). To quantify these findings, we calculated the rank correlation with $\log_2$ fold change in gene expression for unbinned values. $R^2$ values were consistently higher in plots of total mCAC number per gene (including or excluding $[CA]_n$) than for % $[CA]_n$ (Figure 4, middle panels). When the level of CA methylation in $[CA]_n$ was taken into account (Figure 4, right panels), the correlation with transcriptional change did not exceed that of mCAC elsewhere in the genome. Our findings do not support the hypothesis that cytosine modification in $[CA]_n$ has a heightened impact on gene regulation.

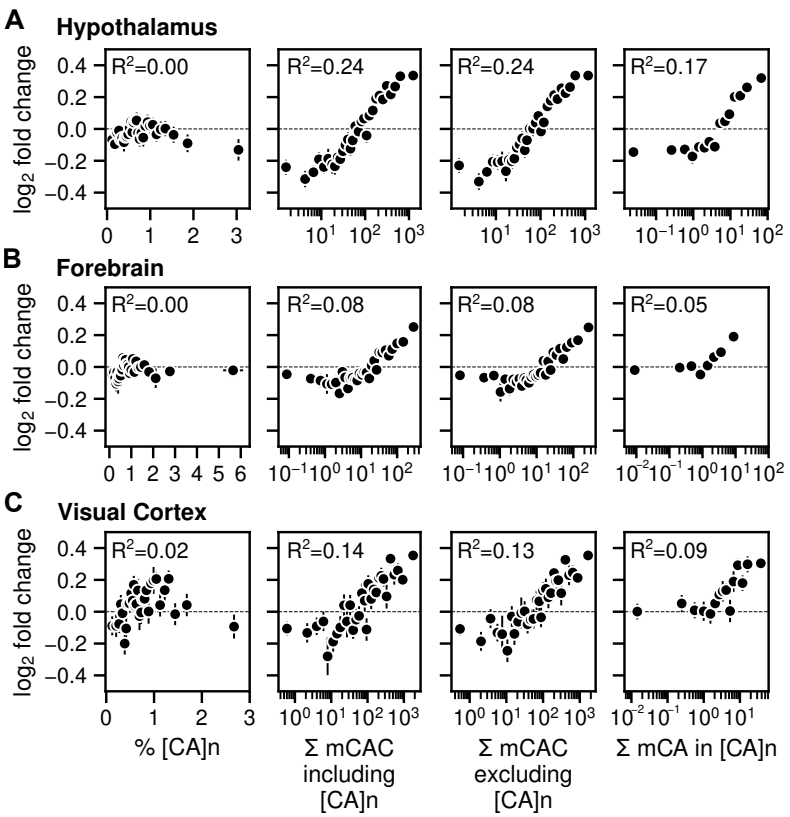



Figure 4 **Gene expression increases in *Mecp2* KO brain regions according to levels of mCAC but does not correlate with the presence of CA repeat arrays. A)** Mean $\log_2$ fold-change in gene expression in *Mecp2* KO versus *Mecp2* WT in mouse hypothalamus (Chen et al., 2015) plotted against the percentage of CA repeats within the gene body (left panels); bins of increasing mCAC number per gene including or excluding CA repeats (Lagger et al., 2017) (centre panels); and bins of increasing mCA per gene exclusively within CA repeats (Lagger et al., 2017) (right panels). Spearman rank correlation ($R^2$) is calculated using unbinned values for every panel. **B)** As described for panel A above but using data for mouse forebrain (Boxer et al., 2020). **C)** As described for panel A above but using data for mouse frontal cortex (Gabel et al., 2015; Lister et al., 2013)

# Discussion

We investigated the possibility that CA repeats are preferred targets of MeCP2 binding due to enrichment of mC or hmC, potentially leading to cooperative binding across the entire tract. In the mouse brain, we detected neither enrichment of CAC modification nor obvious accumulation of bound MeCP2 within $[CA]_n$ repeat clusters. Instead, the data derived from several independent studies indicate that while arrays of $[CA]_n$ do acquire cytosine methylation, the average level of mCAC within them is typical of the surrounding genome. In agreement with this finding, the level of MeCP2 binding within [CA]n repeats is as expected given the frequency of its target motif mCAC. In the absence of MeCP2, gene expression is up-regulated according to mCAC density as reported previously (Guo et al., 2014; Lagger et al., 2017; Cholewa-Waclaw et al., 2019; Chen et al., 2015; Kinde et al., 2016; Boxer et al., 2020), but we found no obvious correlation with the proportion of gene bodies made up of [CA]n repeat blocks unless the frequency of mCAC was taken into account. We conclude that the effect of [CA]n tracts on gene expression depends on the amount of mCAC that they contain.

Several further considerations lead us to question the proposed link between hmC and MeCP2 (Ibrahim et al., 2021). A major reservation concerns the use of antibody detection of hmC, which provided the initial stimulus for the hypothesis of Ibrahim and colleagues (Ibrahim et al., 2021). In their experiments, immunoprecipitation with an antibody directed against hmC revealed prominent apparent peaks of hmC coincident with CA repeat clusters in MEFs. This result is unexpected, as levels of hmC are usually very low in dividing cultured cells compared with neurons. More importantly, others have reported "serious flaws" in the MeDIP method that lead to erroneous reporting of hmC even when both mC and hmC are known to be absent (Lentini et al., 2018). Strikingly, the resulting false positives, which account for 50-99% of regions identified as "enriched" for DNA modifications, are predominantly found at unmodified short repeat arrays, in particular [CA]n. In view of this potentially serious caveat, the evidence for hmC at CA repeats in this MEF cell line must be considered provisional, pending independent biochemical validation.

A second concern relates to the biochemical evidence for the binding specificity of MeCP2. In support of the hypothesis that hmC or mC in CA repeat arrays are the primary targets of MeCP2, Ibrahim et al report that (hmCA)n repeats have a 7-fold higher affinity for MeCP2 in vitro than for the canonical MeCP2 target motif mCG (Ibrahim et al., 2021). However, rather than using symmetrically methylated mCG/mCG, which is a validated target sequence, as a comparator, the authors chose hemi-methylated mCG/CG. The affinity of MeCP2 for hemi-methylated mCG/CG is reproducibly little more than background (Valinluck et al., 2004; Hashimoto et al., 2012; Lei et al., 2019), making this an invalid control. Estimated dissociation constants for the interaction between MeCP2 and symmetrical mCG/mCG are somewhat variable in the literature depending on the details of the assay, ranging from a low of 400nM (Lei et al., 2019) to 15nM (Valinluck et al., 2004) or 10nM (Hashimoto et al., 2012). Notably, these published affinities for mCG/mCG are similar to or higher than the affinity for hmCA-containing repeats reported by Ibrahim et al (410nM).

Finally, the rarity of hmCA in brain is difficult to reconcile with its hypothetical pivotal role. A limitation of many brain methylome datasets is that only bisulfite sequence analysis was performed and therefore mC and hmC are not distinguished. It is clear, however, that while the abundance of neuronal mCA is similar that of mCG in brain, the vast majority of neuronal hmC in confined to hmCG. For example, using a non-destructive method for hmC detection, it was shown that 97.5% of hmC in excitatory neurons is in hmCG, with less than 2.5% in hmCA (Schutsky et al., 2018). This is presumably attributable to the strong preference of Tet enzymes for mCG over mCA as a substrate for mC oxidation (DeNizio et al., 2021; Hu et al., 2015). If levels of hmCA in brain have been systematically underestimated as suggested (Ibrahim et al., 2021), it will be important to provide robust new estimates. As it stands, the moderate affinity of MeCP2 for this ultra-rare motif offers an unlikely basis for comprehensive new models of MeCP2 function.

# Materials and Methods

## Bioinformatic analyses

### Sequencing data sets

Table 1 below details the published data sets used for the analyses. These include Chromatin Immunoprecipitation followed by Sequencing (ChIP-Seq), RNA sequencing (RNA-Seq) and Bisulfite sequencing (WGBS-Seq) libraries from different regions of mouse brain quantifying MeCP2 occupancy, gene expression and DNA methylation levels respectively.

| Brain tissue | Data set | GEO accession |
|---|---|---|
| Hypothalamus | ChIP-seq | GSE66868 (Chen et al., 2015) |
| Hypothalamus | WGBS-seq | GSE84533 (Lagger et al., 2017) |
| Hypothalamus | RNA-seq | GSE66870 (Chen et al., 2015) |
| Forebrain | ChIP-seq | GSE139509 (Boxer et al., 2020) |
| Forebrain | WGBS-seq | GSE128172 (Boxer et al., 2020) |
| Forebrain | RNA-seq | GSE128178 (Boxer et al., 2020) |
| Frontal Cortex | ChIP-seq | GSE67293 (Gabel et al., 2015) |
| Frontal Cortex | WGBS-seq | GSE47966 (Lister et al., 2013) |
| Visual Cortex | RNA-seq | GSE67294 (Gabel et al., 2015) |

Table 1 **Published data sets used for the analyses.**

### RNA-seq analyses

Gene expression analyses were performed for mouse brain RNA-seq data sets. Processed data were downloaded in the form of gene count tables and differential expression analysis was performed using DESeq2. Differentially regulated genes (significance threshold after Benjamini-Hochberg correction p-adjusted value $< 0.05$) in MeCP2 KO and WT mouse brain tissue were sorted according to the total amount of mCAC per gene body excluding [CA]n, binned into 30 equal-sized bins and mean $\log_2$

fold change of each bin is plotted. Error bars represent the standard error of the mean for that bin. Alternatively, the mean $\log_2$ fold change was plotted for differentially regulated genes sorted according to the % CA repeats within the gene body and total methylation within the CA repeats.

### ChIP-seq analyses

Raw fastq reads were downloaded from GEO and subsequent ChIP-seq analysis was performed on mouse genome (mm9) using snakePipes (Bhardwaj et al., 2019) v2.5.1 with the following command line arguments.

```
DNA-mapping --ext .fastq.gz --local -j 12 --snakemakeOptions=-p -j
4 --rerun-incomplete -m mapping --dedup --reads \_1 --GCBias -i
fastq -o dna\_mapping mm9

ChIP-seq -d dna_mapping --singleEnd --fragmentLength 150 --local
--snakemakeOptions -p -j 12 --peakCaller MACS2 --peakCallerOptions
-g mm -m 1 30 -p 0.2 --call-summits --keep-dup all mm9
```

$\log_2$ ChIP-seq signal over Input signal and $\log_2$ Wildtype ChIP-seq signal over MeCP2 Knockout ChIP-Seq signal where available are quantified using bigWigAverageOverBed across genomic locations containing CA repeats, nonCA repeats and rest of the genome.

### WGBS-seq analyses

Processed WGBS-Seq data sets described in Table 1 were downloaded and DNA-methylation ratio (mC basecalls to the count of all reads $C$ as $\hat{m}_C = \frac{m_C}{C}$) at individual nucleoide for CAC, CA and CG contexts. The whole genome is divided into 1 kilobase (kb) windows using bedtools and DNA methylation is calculated as $\hat{M}_C = \sum \hat{m}_C$ within the 1 kb window labelled as mCAC/kb in Fig 2B. For repetitive regions, DNA methylation is calculated as $\hat{M}_C = \sum \hat{m}_C$ within the CA repeat region or nonCA repeat region. Mean DNA methylation level for different DNA sequence contexts is calculated for cytosines with adequate sequence coverage. For $[CA]_n$ and $[CAN]_n$, all cytosines within the genomic loci of respective repeats are considered.

### $[CA]_n$, $[CAN]_n$ and nonCA repeats

The list of genomic locations containing CA and TG repeats was extracted from "Variation and Repeats" group of RepeatMasker track in Mouse (mm9) genome using UCSC table browser functionality [https://genome.ucsc.edu/cgi-bin/hgTables]. For Human (chm13-v1.1), RepeatMasker track was downloaded from processed data (Hoyt et al., 2022). Loci labelled "(CA)n" and "(TG)n" in the RepeatMasker track were used for $[CA]_n$. Loci labelled "(CAA)n", "(TTG)n", "(CAG)n", "(CTG)n", "(CAT)n" and "(ATG)n" were used for $[CAN]_n$. Other repeat sequence loci are considered as nonCA repeats.

### CAC occurrences

After extracting the list of genomic loci for $[CA]_n$, CAC occurrences are calculated using bedtools and jellyfish for $[CA]_n$ and the whole mouse genome.

```
bedtools nuc -fi mm9.fa -bed mm9_CAn_TGn.bed -pattern CAC -C >
    mm9_CAn_TGn.CAC.nuc
bedtools nuc -fi mm9.fa -bed mm9_CAn_TGn.bed -pattern GTG -C >
    mm9_CAn_TGn.GTG.nuc

jellyfish count -t 32 -C -m 3 -s 100M -o mm9_genome.3.jf mm9.fa
```

## Pulldown assay for MeCP2 binding to DNA

This assay was performed as described previously [15] with the following modifications. Biotin-end-labelled double-strand synthetic oligonucleotides (2 g) were coupled to M280-streptavidin Dynabeads according to manufacturer's instructions (Invitrogen). Bead-DNA complex was then co-incubated at 4°C for 1.5 h with nuclear protein (10 g). Nuclear extracts from mouse brain (0.42M salt) were prepared as described [16] and dialysed back into a solution containing 0.15 M NaCl. Following extensive washing, bead-bound proteins were eluted in Laemmli buffer (Sigma) and resolved on a 4–20% SDS-polyacrylamide gel (NEB). The presence of MeCP2 was assayed by western blotting using anti-MeCP2 monoclonal antibody M6818 (Sigma) using IR-dye as a secondary antibody (IRDye 800CW donkey anti-mouse, LI-COR Biosciences). Triplicate assays were scanned then quantified using a LI-COR Odyssey CLx machine and software.

| Name | Sequence of oligonucleotide |
|---|---|
| C1 | 5'-B-tgcgctatgcacttgcgctatgcactttgcgctaatgcacttgcgcttattgcgcacttgcacttttgcacacgcg cacgatgcgcttaatgcgcgattgcacacgctgcacacacgcgctttgca-3' |
| $[CA]_7$ | 5'-B-tgcgctatgcacttgcgctatgcactttgcgctaatgcacttgcgCACACACACACACActtattgcgcacttgca cttttgcacacgcgcacgatgcgcttaatgcgcgattgcacacgctgcacacacgcgctttgca-3' |
| $[CA]_7$-1mC | 5'-B-tgcgctatgcacttgcgctatgcactttgcgctaatgcacttgcgCACAmCACACACACActtattgcgcacttgca cttttgcacacgcgcacgatgcgcttaatgcgcgattgcacacgctgcacacacgcgctttgca-3' |
| $[CA]_7$-4mC | 5'-B-tgcgctatgcacttgcgctatgcactttgcgctaatgcacttgcgmCACAmCACAmCACAmCActtattgcgcacttg cacttttgcacacgcgcacgatgcgcttaatgcgcgattgcacacgctgcacacacgcgctttgca-3' |
| C2 | 5'-B-tgcgctatgcacttgcgctatgcactttgcgctaatgcacttgcgCAGACAGACAGACActtattgcgca cttgcacttttgcacacgcgcacgatgcgcttaatgcgcgattgcacacgctgcacacacgcgctttgca-3' |
| CAC | 5'-B-cgcactttgcactatgcacttgcactatgcactttgcactaatgcacttgcacttattgcacacttgca cttttgcacacacgcacgatgcacttaatgcacgattgcacacactgcacacacgcactttgcacactgca-3' |
| mCAC | 5'-B-cgcactttgmCACtatgcacttgmCACtatgcactttgmCACtaatgcacttgmCACttattgmCACacttgcac ttttgcacamCACgcacgatgmCACttaatgmCACgattgcacamCACtgcacacacgmCACtttgcacactgca-3' |
| CG | 5'-B-cgcactttgCGctatgcacttgCGctatgcactttgCGctaatgcacttgCGcttattgCGcacttgcacttttg cacaCGcgcacgatgCGcttaatgCGcgattgcacaCGctgcacacacgCGctttgcacactgca-3' |
| mCG | 5'-B-cgcactttgmCGctatgcacttgmCGctatgcactttgmCGctaatgcacttgmCGgcttattgmCGcacttgcac ttttgcacamCGcgcacgatgmCGcttaatgmCGcgattgcacamCGctgcacacacgmCGctttgcacactgca-3' |

Table 2 **Oligonucleotide sequences for the probes used in the pulldown assay.** B=biotin, m=methyl group. All molecules were annealed to the appropriate methylated or non-methylated reverse oligonucleotide.

# Acknowledgements

We thank Raphaël Pantier and Matthew Lyst for providing critical feedback on the manuscript. This work was supported by Wellcome Centre grant to A.B., who is also a member of the Simons Initiative for the Developing Brain. K. C. was supported by a scholarship from College of Science and Engineering, University of Edinburgh.

# Author Contributions

Conceptualization, A.B., K.C. and J.C.; Methodology, K.C. and J.C.; Software, K.C.,S.W.; Formal Analysis, K.C.,S.W.; Investigation, K.C., J.C. and S.W.; Writing – Original Draft, K.C., J.C. and A.B.; Writing – Review & Editing, K.C., J.C., S.W., S.K. and A.B.; Supervision, A.B; Funding acquisition, A.B.

# References


P. Ayata. Decoding 5HMC as an Active Chromatin Mark in the Brain and its Link to Rett Syndrome. *Student Theses and Dissertations*, Jan. 2013.

V. Bhardwaj, S. Heyne, K. Sikora, L. Rabbani, M. Rauer, F. Kilpert, A. S. Richter, D. P. Ryan, and T. Manke. snakePipes: Facilitating flexible, scalable and integrative epigenomic analysis. *Bioinformatics*, 35(22):4757–4759, Nov. 2019. ISSN 1367-4803. doi:10.1093/bioinformatics/btz436.

L. D. Boxer, W. Renthal, A. W. Greben, T. Whitwam, A. Silberfeld, H. Stroud, E. Li, M. G. Yang, B. Kinde, E. C. Griffith, B. Bonev, and M. E. Greenberg. MeCP2 Represses the Rate of Transcriptional Initiation of Highly Methylated Long Genes. *Molecular Cell*, 77(2):294–309.e9, Jan. 2020. ISSN 1097-2765. doi:10.1016/j.molcel.2019.10.032.

L. Chen, K. Chen, L. A. Lavery, S. A. Baker, C. A. Shaw, W. Li, and H. Y. Zoghbi. MeCP2 binds to non-CG methylated DNA as neurons mature, influencing transcription and the timing of onset for Rett syndrome. *Proceedings of the National Academy of Sciences*, 112(17):5509–5514, Apr. 2015. ISSN 0027-8424, 1091-6490. doi:10.1073/pnas.1505909112.

J. Cholewa-Waclaw, R. Shah, S. Webb, K. Chhatbar, B. Ramsahoye, O. Pusch, M. Yu, P. Greulich, B. Waclaw, and A. P. Bird. Quantitative modelling predicts the impact of DNA methylation on RNA polymerase II traffic. *Proceedings of the National Academy of Sciences*, 116(30):14995–15000, July 2019. ISSN 0027-8424, 1091-6490. doi:10.1073/pnas.1903549116.

J. C. Connelly, J. Cholewa-Waclaw, S. Webb, V. Steccanella, B. Waclaw, and A. Bird. Absence of MeCP2 binding to non-methylated GT-rich sequences in vivo. *Nucleic Acids Research*, 48(7):3542–3552, Apr. 2020. ISSN 0305-1048. doi:10.1093/nar/gkaa102.

A. de Mendoza, D. Poppe, S. Buckberry, J. Pflueger, C. B. Albertin, T. Daish, S. Bertrand, E. de la Calle-Mustienes, J. L. Gómez-Skarmeta, J. R. Nery, J. R. Ecker, B. Baer, C. W. Ragsdale, F. Grützner, H. Escriva, B. Venkatesh, O. Bogdanovic, and R. Lister. The emergence of the brain non-CpG methylation system in vertebrates. *Nature Ecology & Evolution*, pages 1–10, Jan. 2021. ISSN 2397-334X. doi:10.1038/s41559-020-01371-2.

J. E. DeNizio, B. J. Dow, J. C. Serrano, U. Ghanty, A. C. Drohat, and R. M. Kohli. TET-TDG Active DNA Demethylation at CpG and Non-CpG Sites. *Journal of Molecular Biology*, 433(8):166877, Apr. 2021. ISSN 0022-2836. doi:10.1016/j.jmb.2021.166877.

H. W. Gabel, B. Kinde, H. Stroud, C. S. Gilbert, D. A. Harmin, N. R. Kastan, M. Hemberg, D. H. Ebert, and M. E. Greenberg. Disruption of DNA-methylation-dependent long gene repression in Rett syndrome. *Nature*, 522(7554):89–93, June 2015. ISSN 0028-0836, 1476-4687. doi:10.1038/nature14319.

J. U. Guo, Y. Su, J. H. Shin, J. Shin, H. Li, B. Xie, C. Zhong, S. Hu, T. Le, G. Fan, H. Zhu, Q. Chang, Y. Gao, G.-l. Ming, and H. Song. Distribution, recognition and regulation of non-CpG methylation in the adult mammalian brain. *Nature Neuroscience*, 17(2): 215–222, Feb. 2014. ISSN 1546-1726. doi:10.1038/nn.3607.

H. Hashimoto, Y. Liu, A. K. Upadhyay, Y. Chang, S. B. Howerton, P. M. Vertino, X. Zhang, and X. Cheng. Recognition and potential mechanisms for replication and erasure of cytosine hydroxymethylation. *Nucleic Acids Research*, 40(11):4841–4849, June 2012. ISSN 0305-1048. doi:10.1093/nar/gks155.

S. J. Hoyt, J. M. Storer, G. A. Hartley, P. G. Grady, A. Gershman, C. Limouse, R. Halabian, L. Wojenski, and R. J. O'Neill. From telomere to telomere: the transcriptional and epigenetic state of human repeat elements analysis code: T2T-CHM13, Jan. 2022. URL https://doi.org/10.5281/zenodo.5895031.

L. Hu, J. Lu, J. Cheng, Q. Rao, Z. Li, H. Hou, Z. Lou, L. Zhang, W. Li, W. Gong, M. Liu, C. Sun, X. Yin, J. Li, X. Tan, P. Wang, Y. Wang, D. Fang, Q. Cui, P. Yang, C. He, H. Jiang, C. Luo, and Y. Xu. Structural insight into substrate preference for TET-mediated oxidation. *Nature*, 527(7576):118–122, Nov. 2015. ISSN 1476-4687. doi:10.1038/nature15713.

A. Ibrahim, C. Papin, K. Mohideen-Abdul, S. L. Gras, I. Stoll, C. Bronner, S. Dimitrov, B. P. Klaholz, and A. Hamiche. MeCP2 is a microsatellite binding protein that protects CA repeats from nucleosome invasion. *Science*, 372(6549), June 2021. ISSN 0036-8075, 1095-9203. doi:10.1126/science.abd5581.

B. Kinde, D. Y. Wu, M. E. Greenberg, and H. W. Gabel. DNA methylation in the gene body influences MeCP2-mediated gene repression. *Proceedings of the National Academy of Sciences*, 113(52):15114–15119, Dec. 2016. ISSN 0027-8424, 1091-6490. doi:10.1073/pnas.1618737114.

S. Lagger, J. C. Connelly, G. Schweikert, S. Webb, J. Selfridge, B. H. Ramsahoye, M. Yu, C. He, G. Sanguinetti, L. C. Sowers, M. D. Walkinshaw, and A. Bird. MeCP2 recognizes cytosine methylated tri-nucleotide and di-nucleotide sequences to tune transcription in the mammalian brain. *PLOS Genetics*, 13(5):e1006793, May 2017. ISSN 1553-7404. doi:10.1371/journal.pgen.1006793.

L. A. Lavery, K. Ure, Y.-W. Wan, C. Luo, A. J. Trostle, W. Wang, H. Jin, J. Lopez, J. Lucero, M. A. Durham, R. Castanon, J. R. Nery, Z. Liu, M. Goodell, J. R. Ecker, M. M. Behrens, and H. Y. Zoghbi. Losing Dnmt3a dependent methylation in inhibitory neurons impairs neural function by a mechanism impacting Rett syndrome. *eLife*, 9:e52981, Mar. 2020. ISSN 2050-084X. doi:10.7554/eLife.52981.

M. Lei, W. Tempel, S. Chen, K. Liu, and J. Min. Plasticity at the DNA recognition site of the MeCP2 mCG-binding domain. *Biochimica et Biophysica Acta (BBA) - Gene Regulatory Mechanisms*, 1862(9):194409, Sept. 2019. ISSN 1874-9399. doi:10.1016/j.bbagrm.2019.194409.

A. Lentini, C. Lagerwall, S. Vikingsson, H. K. Mjoseng, K. Douvlataniotis, H. Vogt, H. Green, R. R. Meehan, M. Benson, and C. E. Nestor. A reassessment of DNA-immunoprecipitation-based genomic profiling. *Nature Methods*, 15(7):499–504, July 2018. ISSN 1548-7105. doi:10.1038/s41592-018-0038-7.

J. D. Lewis, R. R. Meehan, W. J. Henzel, I. Maurer-Fogy, P. Jeppesen, F. Klein, and A. Bird. Purification, sequence, and cellular localization of a novel chromosomal protein that binds to Methylated DNA. *Cell*, 69(6):905–914, June 1992. ISSN 0092-8674. doi:10.1016/0092-8674(92)90610-O.

R. Lister, E. A. Mukamel, J. R. Nery, M. Urich, C. A. Puddifoot, N. D. Johnson, J. Lucero, Y. Huang, A. J. Dwork, M. D. Schultz, M. Yu, J. Tonti-Filippini, H. Heyn, S. Hu, J. C. Wu, A. Rao, M. Esteller, C. He, F. G. Haghighi, T. J. Sejnowski, M. M. Behrens, and J. R. Ecker. Global Epigenomic Reconfiguration During Mammalian Brain Development. *Science*, 341(6146):1237905, Aug. 2013. ISSN 0036-8075, 1095-9203. doi:10.1126/science.1237905.

E. K. Schutsky, J. E. DeNizio, P. Hu, M. Y. Liu, C. S. Nabel, E. B. Fabyanic, Y. Hwang, F. D. Bushman, H. Wu, and R. M. Kohli. Nondestructive, base-resolution sequencing of 5-hydroxymethylcytosine using a DNA deaminase. *Nature Biotechnology*, 36(11):1083–1090, Nov. 2018. ISSN 1546-1696. doi:10.1038/nbt.4204.

R. Tillotson, J. Cholewa-Waclaw, K. Chhatbar, J. C. Connelly, S. A. Kirschner, S. Webb, M. V. Koerner, J. Selfridge, D. A. Kelly, D. D. Sousa, K. Brown, M. J. Lyst, S. Kriaucionis, and A. Bird. Neuronal non-CG methylation is an essential target for MeCP2 function. *Molecular Cell*, 81(6):1260–1275.e12, Mar. 2021. ISSN 1097-2765. doi:10.1016/j.molcel.2021.01.011.

V. Valinluck, H.-H. Tsai, D. K. Rogstad, A. Burdzy, A. Bird, and L. C. Sowers. Oxidative damage to methyl-CpG sequences inhibits the binding of the methyl-CpG binding domain (MBD) of methyl-CpG binding protein 2 (MeCP2). *Nucleic Acids Research*, 32(14): 4100–4108, July 2004. ISSN 0305-1048. doi:10.1093/nar/gkh739.

J. M. Weitzel, H. Buhrmester, and W. H. Strätling. Chicken MAR-binding protein ARBP is homologous to rat methyl-CpG-binding protein MeCP2. *Molecular and Cellular Biology*, 17(9):5656–5666, Sept. 1997. doi:10.1128/MCB.17.9.5656.