Mathematical modelling reveals how MeCP2 restrains transcriptional elongation in human neurons

Justyna Cholewa-Waclaw[1,8], Ruth Shah[1,8], Shaun Webb[1], Kashyap Chhatbar[1], Bernard Ramsahoye[2], Oliver Pusch[3], Miao Yu[4], Philip Greulich[5,6], Bartlomiej Waclaw[7,9], and Adrian Bird[1,9]

[1] The Wellcome Centre for Cell Biology, The King's Buildings, University of Edinburgh, Max Born Crescent, Edinburgh, EH9 3BF, UK;
[2] Institute of Genetics and Molecular Medicine, University of Edinburgh, Western General Hospital Campus, Crewe Road, Edinburgh, EH4 2XU, UK;
[3] Center for Anatomy and Cell Biology, Medical University of Vienna, Schwarzspanierstrasse 17, Vienna, 1090, Austria;
[4] Ludwig Institute for Cancer Research, La Jolla, California, 92093, USA;
[5] Mathematical Sciences, University of Southampton, Southampton, SO17 1BJ, UK;
[6] Institute for Life Sciences, University of Southampton, Southampton, SO17 1BJ, UK;
[7] School of Physics and Astronomy, University of Edinburgh, James Clerk Maxwell Building, Peter Guthrie Tait Road, Edinburgh, EH9 3FD, UK;
[8] Equal contributions;
[9] Corresponding authors: Adrian Bird, a.bird@ed.ac.uk; Bartlomiej Waclaw, bwaclaw@ph.ed.ac.uk;

ABSTRACT

Gene expression patterns depend on the interaction of diverse transcription factors with their target genes. While many factors have a restricted number of targets, some appear to affect transcription globally. An example of the latter is MeCP2; an abundant chromatin-associated protein that is mutated in the neurological disorder Rett Syndrome. To understand how MeCP2 affects transcription, we integrated mathematical modelling with quantitative experimental analysis of human neurons expressing graded levels of MeCP2. We first used a model of MeCP2-DNA binding to demonstrate that changes in gene expression reflect MeCP2 density downstream of transcription initiation. We then tested five biologically plausible hypotheses for the effect of MeCP2 on transcription. The only model compatible with the data involved slowing down of RNA polymerase II by MeCP2, causing reduced transcript output due to polymerase queueing. Our general approach may prove fruitful in deciphering the mechanisms by which other global regulators choreograph gene expression.

INTRODUCTION

Gene expression is a complex process affected by many factors. Traditionally, interactions between genes have been visualized as a gene regulatory network in which transcription factors are classified either as repressors or activators, each with a few well-defined target genes. This view becomes less informative when a factor interacts with the majority of genes. Such global regulators of transcription may either interact with the transcriptional machinery

itself, for example by directly affecting the initiation rate of Pol II, or bind to abundant DNA features that are widespread in the genome, leading to altered transcriptional initiation or elongation. The phenomenon of global transcriptional regulation is well known in bacteria. In *Escherichia coli*, for example, seven global regulators directly modulate expression of half of all the genes which are responsible for cellular metabolism (Perrenoud and Sauer, 2005).

Many eukaryotic chromatin-binding factors can potentially act as global modulators, including "typical" transcription factors such as c-Myc, PLZF and SALL4 that bind thousands of sites in gene promoters (Lin et al., 2012; Lovelace et al., 2016). In principle, binding need not be restricted to promoters or even specific DNA sequences, as any protein that interacts broadly with DNA may modulate gene expression at a global level. For example, histones that form the core of nucleosomes in eukaryotes are known to directly block transcription (Li et al., 2007; Workman and Kingston, 1998). To facilitate transcription, DNA must thus be unwound from the histone core, and histone modifications such as acetylation and methylation affect this process in various ways (Berry et al., 2017; Ouararhni et al., 2006; Woo et al., 2017). Although the precise mechanisms by which global regulation is achieved are still unclear, most hypotheses rely on alteration of the initiation rate of transcription. For example, H3K27 histone methylation has been suggested to affect chromatin compaction near the promoter (Berry et al., 2017). Other studies have considered the presence of dynamic promoter-proximal obstacles (Hao et al., 2017), or interactions between RNA polymerases (Epshtein, 2003; Sneppen et al., 2005).

By comparison, global regulation via processes that occur in the bodies of genes rather than at promoters has received less attention. Potential examples include chromatin compaction mediated through histone modifications or other features. In particular, DNA methylation is known to affect gene expression. This has been well-studied in the context of promoter methylation, which represses transcription. However, DNA methylation also seems to affect transcription via gene bodies; RNA Pol II has been observed to slow down on exons that are more methylated than the surrounding introns (Jonkers and Lis, 2015). Since methylated DNA is targeted by a broad family of methyl-CpG binding proteins (Hendrich and Bird, 1998), the presence of these proteins in gene bodies may have a direct effect on transcription. In this work we focus on the putative global modulator, methyl CpG binding protein 2 (MeCP2), which is highly expressed in adult human brain, especially in neurons (Lewis et al., 1992b; Skene et al., 2010). Loss-of-function mutations in the X-linked *MECP2* gene or excessive amounts of MeCP2 caused by duplication of the gene lead, respectively, to the severe neurological disorders Rett Syndrome (Amir et al., 1999) and MeCP2 duplication syndrome (Ramocki et al., 2010). Missense mutations responsible for Rett syndrome accumulate mainly in two regions of MeCP2, coincident with the Methyl-CpG Binding Domain (MBD), whose binding to DNA requires methylated cytosine (Ho et al., 2008; Nan et al., 1993; 1996), and a downstream NCoR interaction domain (NID), which binds the TBL1/TBLR1 subunits of the NCoR co-repressor complex (Kruusvee et al., 2017; Lyst et al., 2013). In addition to mCpG, MeCP2 also binds to methylated CpA, predominantly in the trinucleotide mCAC, whose accumulation in the brain coincides with increased neuronal MeCP2 expression (Gabel et al., 2015; Guo et al., 2014; Lagger et al., 2017; Lister et al., 2013a).

Although the interactions between MeCP2 and DNA are well established, its precise role in gene expression is less clear. It is evident that MeCP2 can repress transcription in a DNA methylation-dependent manner (Guy et al., 2001; Nan et al., 1997) and that repression is

reduced in MeCP2 mutants that cannot bind co-repressor complexes or by the addition of HDAC inhibitors (Bird et al., 1998; Lyst et al., 2013; Nan et al., 1997; 1998). In mouse brain it is reported that depletion of MeCP2 increases global histone acetylation (Shahbazian, 2002; Skene et al., 2010) and modestly elevates expression of repeats and retrotransposons (Skene et al., 2010). However, MeCP2 does not behave as a conventional transcription factor with a few well-defined targets, as its binding site occurs on average every ~100 base pairs. Previous RNA-seq data, predominantly from mouse brain regions over-expressing or deficient in MeCP2, indicate that the magnitude of effects on transcription are often small. However, correlations between MeCP2 occupancy across genes and the degree of transcriptional change when MeCP2 is removed or over-expressed have been reported, raising questions about the regulatory processes at work (Gabel et al, 2015; Kinde et al, 2016; Lagger et al, 2017).

Here we set out to resolve the mechanism of MeCP2-dependent transcriptional regulation. Because MeCP2 binding sites occur in the vast majority of genes, we reasoned that most are likely to be influenced to some extent by its presence. To confront the technical and analytical challenges posed by modest changes in expression of large numbers of genes, we adopted a quantitative approach that combined deep, high quality datasets obtained from a uniform population of Lund Human Mesencephalic (LUHMES)-derived human dopaminergic neurons (Scholz et al., 2011) with mathematical modelling. We created a spectrum of LUHMES cell lines expressing distinct levels of MeCP2. Using transposase-accessible chromatin sequencing (ATAC-seq) and chromatin immunoprecipitation (ChIP-seq) together with mathematical modelling, we detected a robust footprint of MeCP2 binding to mCG and mCA *in vivo* and determined the amount of MeCP2 bound to DNA. Quantification of mRNA abundance by RNA-seq revealed a previously unknown relationship between changes in transcription and the density of mCG on gene bodies. To explain this observation, we proposed and tested several distinct mechanistic models. The only model consistent with our experimental results is one in which MeCP2 leads to slowing down of RNA polymerase II progression through a transcription unit. Importantly, mutant MeCP2 that is unable to bind NCoR fails to repress efficiently, suggesting that repression depends upon this interaction.

RESULTS

Our approach is illustrated in Figures 1A and 1B. We first created cell lines with varying levels of MeCP2 (Figure 1A) and then used a range of techniques to determine: DNA methylation (whole-genome bisulfite sequencing - WGBS), total number of MeCP2 molecules in each cell line, MeCP2 chromatin occupancy (ChIP-seq, ATAC-seq), changes in transcription (RNA-seq) and chromatin accessibility (ATAC-seq). We next constructed mathematical models of MeCP2-chromatin binding; fitting the models to experimental data enabled us to precisely determine the density of MeCP2 on the chromatin. These results were subsequently used in mathematical models of MeCP2 transcriptional regulation. A comparison between the models and experiments enables us to falsify candidate models (Figure 1B).

Seven LUHMES-derived cell lines expressing a wide range of MeCP2 levels were created and assayed by western blotting (Figure 1C). Based on morphology, expression of neuronal markers and the temporal profile of MeCP2 protein during differentiation (Figures S1A–D), we chose to analyse neuronal preparations consistently on day 9. The following cell lines

expressing varying MeCP2 levels were generated: zero MeCP2 (KO; purple) (Shah et al., 2016), low levels via two independent knock-downs (KDs; dark blue, 10-20% WT and lighter blue, 30-40% WT), wild-type (WT, green), and three over-expression levels (OE; yellow, 3-fold; orange, 4-fold; and red, 11-fold; see Table S1). The resulting MeCP2 levels corresponded to approximately 2.5×10^6 MeCP2 molecules per nucleus in WT (Figure S1E) and 2.7×10^7 molecules in OE 11x (Figure 1D). For comparison, MeCP2 abundance in our highest expressing cell lines exceeds that in mature neurons derived from mouse brain (~1.5 x 10^7 molecules per nucleus; (Skene et al., 2010)). Immunostaining for MeCP2 showed uniform expression of MeCP2 in most cell lines, with the exception of lines KD2 and OE 3x, which displayed somewhat variable levels (Figure 1E; middle panel). All cell lines differentiated with similar kinetics (Figures S1J and S1M), expressing similar levels of cellular markers (Figures S1K and S1N) and acquiring similar neurite length (Figures S1L and S1O). Visualization of neurofilament (NF) staining and other neuron-specific markers indicated that the differing levels of MeCP2 do not significantly affect neuronal differentiation in this system (Figures 1E and S1N). These findings agree with previous observations that *Mecp2*-null neurons differentiate efficiently *in vitro* (Kishi and Macklis, 2004) and *Mecp2*-null mice show no overt phenotypic defects until ~4 weeks of age (Chen et al., 2001; Guy et al., 2001).

Methylated CG is the dominant DNA modification in LUHMES neurons

MeCP2 was initially detected as a m5C binding protein and subsequent work has confirmed this property (Baubec et al., 2013; Gabel et al., 2015; Lagger et al., 2017; Lewis et al., 1992a; Skene et al., 2010). As a reader of this epigenetic mark, MeCP2 is not expected to affect the level of DNA methylation. Accordingly, global levels of DNA methylation were consistent in all cell lines tested, with ~3.7% of all cytosines being methylated (Figure S1P). We determined the detailed pattern of DNA methylation and hydroxymethylation at single base pair resolution in WT cells using whole-genome bisulfite sequencing, TAB-seq and ox-BS. The analysis revealed that 70% of CGs in the genome are methylated, equivalent to ~40 million mCGs per haploid genome (Figure 1F). Notably, only 2.1% of 5mC is in the CA context, corresponding to ~1 million mCA moieties per haploid genome (Figures 1F and G), as compared with up to 50% 5mC in mCA in NeuN-positive adult human neurons (Lister et al., 2013b). Additionally, only ~2 million CGs are hydroxymethylated in the LUHMES haploid genome (Figure 1F). In summary, methylation of LUHMES neuronal genomes occurs predominantly in the mCG context, with much lower levels of hmCG and non-CG methylation.

MeCP2 binds predominantly methylated CG genome-wide

To map the binding of MeCP2 in these human neurons, we performed quantitative MeCP2 ChIP-seq for KO, WT, OE 4x and OE 11x, and simultaneously developed a computer model that simulates the ChIP-seq procedure and MeCP2 binding *in vivo* (Figure 2A). A constant amount of *Drosophila* chromatin was added as an internal standard (Figure S2A) and we confirmed that ChIP enrichment (human DNA relative to *Drosophila* DNA) was proportional to the level of MeCP2 in each cell line (Figures 2B and S2B). To determine the distribution of MeCP2 in the genome we accumulated the number of MeCP2 ChIP-seq reads on and around certain genomic features, including methylated/unmethylated dinucleotides, CpG islands and promoters. When ChIP enrichment profiles were centred at methylated CGs, a peak whose height increases with the level of MeCP2 was visible (Figure 2C). However, the peak height was not proportional to the amount of MeCP2 in the cell, and a peak was also present in KO

cells indicating a bias unrelated to MeCP2. To interpret this peak and to extract MeCP2 occupancy (fraction of mCG occupied by MeCP2), we fitted the enrichment profiles obtained from the *in silico* generated data to the experimental profiles and extracted the absolute MeCP2 density. Based on good agreement between the simulated and experimental profiles (Figure 2C), we determined that <10% of mCG sites are occupied by MeCP2 in OE 11x genome (Figure S2E), suggesting that fewer than 8 million mCGs are bound by MeCP2 despite the ~25 million MeCP2 molecules present in the nucleus. The specificity of this approach was confirmed by deriving MeCP2 binding profiles over unmethylated CGs, resulting in a significant drop in enrichment (Figure 2D). The simulation also reproduces a lack of enrichment over GT dinucleotides (Figure S2F, upper panel), and, despite its relative rarity, an enrichment over mCA (Figure S2F, bottom panel). A model in which MeCP2 binds only to mCG (no mCA binding) does not reproduce the mCA peak (Figure S2G, bottom panel), further supporting specific binding *in vivo* to mCG and mCA. A weak peak found in MeCP2 KO ChIP-seq over mCG and mCA is also reproduced by the model in the absence of MeCP2 (Figure 2C, KO panel). Its MeCP2-independent origin is unknown, but may be due to biases in sonication, PCR amplification or sequencing.

To further assess the binding model, we checked if it could correctly predict other features of the ChIP-seq data. Figure 2E shows very good agreement between experimental and simulated ChIP enrichment plotted as a function of mCG density. This implies a linear relationship between MeCP2 density on the DNA and mCG density. Figure 2F shows that enrichment profiles in regions near the transcription start site (TSS) are also correctly reproduced by the model, confirming depletion of MeCP2 from promoters. In addition, a significant depletion of MeCP2 over unmethylated CGIs is visible by inspection of individual tracks (Figure S2H) and a dip in both the experimental and simulated data corresponds to reduced binding in CGIs (Figure S2I). Figure 2G summarizes model-based quantification of MeCP2 occupancy on various genomic features. The large difference in MeCP2 occupancy between methylated and unmethylated CGIs and the moderate binding in the regions of gene bodies, intergenic regions and promoters agrees with the average methylation status of each feature.

In vivo MeCP2 footprint is visible on methylated CG and CA

To derive an independent measure of absolute MeCP2 density on the DNA and to detect its molecular footprint with high resolution, we performed ATAC-seq (Buenrostro et al., 2013) in KO, WT, OE 4x and OE 11x neurons. Transposase Tn5 cuts the DNA in regions devoid of DNA-bound proteins, hence the number of Tn5 “insertions” obtained from ATAC-seq represents DNA accessibility within chromatin. To interpret the data, we developed a computer model of MeCP2-chromatin binding and the molecular consequences of the ATAC-seq procedure (Figure 3A). We first obtained Tn5 insertion profiles around the same genomic features as for ChIP-seq. The profiles showed multiple peaks and valleys surrounding the mCG dinucleotide (Figure S3A). As this structure also occurs in KO neurons we concluded that it is related to factors other than MeCP2, such as the insertion bias of the transposase (Figure S3B,C). The GC bias of sequencing (Figure S3D) and time of incubation with transposase (Figure S3E) also affect the profiles. To remove these biases, we calculated relative insertion profiles between all tested cell lines and KO1. This resulted in a clear signal whose amplitude increased with MeCP2 concentration (Figure 3B). The same analysis performed for non-methylated CG showed no equivalent profile (Figure 3C). We interpret this

"molecular footprint" as a direct evidence of MeCP2 binding to its cognate recognition sequence. Excellent agreement between simulated and experimental footprints (Figure 3B) allowed us to draw several conclusions. First, the size of the footprint is consistent with a protein that obscures ~11bp, in agreement with *in vitro* MeCP2-DNA binding analysis (Nan et al., 1993; Nikitina et al., 2007). Second, the amplitude of the footprint is linear with respect to MeCP2 concentration (Figure 3D). Third, in agreement with the upper estimate of bound MeCP2 deduced from MeCP2 ChIP-seq (<10% for OE 11x), only 6.3% of mCG sites are actually occupied by MeCP2 in OE 11x neurons, with less than 1% occupancy in WT (Figure 3D). Considering the number of MeCP2 molecules in OE 11x neurons ($\sim 2.5 \times 10^7$) and the total number of mCG sites (~8×10^7/diploid genome), we calculate that only $(p = 0.063) \times (8 \times 10^7)/(2.5 \times 10^7) \approx 20\%$ of MeCP2 molecules are bound to DNA at any given time in OE 11x. The same methodology allowed us to detect a footprint of MeCP2 binding to mCA but no footprint at control GT (unmethylated) or TA dinucleotides (Figure S3F).

Overexpressing MeCP2 decreases accessibility of promoters and gene bodies

We used the ATAC-seq data to infer the effect of MeCP2 on chromatin accessibility. We first identified regions of high and low ("background") ATAC-seq insertion probabilities and defined accessibility a as the total number of insertions in the peaks divided by the average insertion counts. The value of a is proportional to the fraction of cells in which a given region is accessible to Tn5. Figure 3E shows frequency distributions of accessibility a in gene bodies for different cell lines. A striking decrease in accessibility *a* is evident between the two OE cell lines compared with KO1, while two independent KO cell lines (KO1 and KO2) are indistinguishable from each other. A similar effect is also detected at promoters (Figure S3G). The mean accessibility of genomic DNA decreases monotonously as MeCP2 abundance increases (Figure 3F). Superficially, reduced accessibility for DNA binding proteins, such as transcription factors, might be expected to cause a negative constraint on transcription; we test this and other models below.

Global changes in transcription correlate with MeCP2 occupancy and mCG density in transcription units

To determine the influence of MeCP2 on transcription, we performed RNA-seq on all seven cell lines expressing varying levels of MeCP2 (Table S1), and validated the results by qPCR (Figures S4A,B). We also measured total RNA levels and found that it was modestly reduced in KO neurons (~8% less compared with control), as seen previously in cultured human and mouse neurons and mouse brain (Figure S4C) (Lagger et al., 2017; Li et al., 2013; Yazdani et al., 2012). For each gene represented in the RNA-seq data, we calculated the Log2-fold change relative to the appropriate control (Log2FC) expressed as transcript abundance (transcripts per million, TPM; Figure 4A). An initial analysis that was restricted to "significantly-changed" genes revealed a strong reciprocal relationship of Log2FC to mCG density ρ_{mCG} (Figure S4D). However, since RNA-seq only measures expression relative to some common average, rather than absolute expression, there is uncertainty in defining the basal level (Log2FC=0) and significance (p-values). In view of this limitation, we considered reliance on arbitrarily determined "significance" unsuitable for mathematical modelling and therefore adopted a different approach. We binned all expressed genes with sufficient coverage in WGBS, regardless of their Log2FC ($\sim 15{,}000$ genes) according to ρ_{mCG} and calculated the

average Log2FC in each bin. This procedure reduced noise and removed the effect of gene-gene interactions, enabling us to detect small changes in expression which might otherwise be obscured by other regulatory mechanisms and statistical noise. Figure 4B shows that most genes have mCG densities between 0.8 and 4 mCG per 100bp. These genes exhibit a strong, reciprocal relationship between methylation in gene bodies and Log2FC, dependent on whether MeCP2 was overexpressed or deficient (Figure 4C). In KO and KD neurons, Log2FC relative to WT increased with ρ_{mCG} in gene bodies (Figure 4C, upper panel), whereas in OE 3x, 4x and 11x neurons, increasing ρ_{mCG} correlated with lowered expression (Figure 4C, bottom panel). Previous transfection-based assays of MeCP2-mediated repression also detected an initial steep decrease in expression that plateaued at higher mCG densities (Nan et al., 1997), mimicking MeCP2-overexpressing neurons in this analysis.

To determine the relationship between the change in gene expression and the concentration of MeCP2 binding sites, we calculated the maximum slope (Methods) of the Log2FC in expression versus ρ_{mCG} in gene bodies (Figure 4C, black lines). The slope is strikingly proportional to MeCP2 density in gene bodies (Figure 4D). In contrast, when this analysis was repeated for promoter regions rather than gene bodies the slope was close to zero, indicating only a weak dependence on ρ_{mCG} in promoters (Figure 4E,F). Thus, although chromatin accessibility over promoters is constrained in MeCP2 OE cell lines (Figure S3G), promoter methylation appears to have little impact on the expression of genes that are actively transcribed. Since we previously established (ChIP-seq, ATAC-seq) that ρ_{mCG} is a strong predictor of MeCP2 density on the chromatin (Figures 4C,E), we infer that MeCP2 density in gene bodies has a strong effect on transcription, in agreement with a previous proposal (Kinde et al., 2016).

The results suggest that MeCP2 globally modulates transcription, however, they do not distinguish whether MeCP2 acts as a repressor or an activator of transcription as genes both increase and decrease their expression level in OE and KO/KD lines. Moreover, the apparent correlation between expression and mCG density does not prove causality, as Log2FC correlates with other quantities, including gene length, CG density regardless of methylation, and even GC content (Figures S4E-G).

Mathematical modelling excludes three plausible models of MeCP2 transcriptional regulation

To gain mechanistic insight into MeCP2 transcriptional regulation we proposed a range of models, derived predictions from each model, and compared these predictions with the experimental data to eliminate those models that are incompatible with the data. All five models are based on a commonly accepted paradigm of gene expression (Figure 5A), well supported in both eukaryotic and bacterial cells (Angel et al., 2015; Arjun Raj, 2008; Shahrezaei et al., 2008). A gene *i* switches between two states, ON or OFF, with rates k_{ON}, k_{OFF}. Expression is negligible in the OFF state, whereas in the ON state it assumes rate J_i (transcripts/second). The rate J depends on the initiation rate α and elongation rate v. Transcripts are degraded with rate d_i. We define f_i as the fraction of cells that at any given time have gene i in the ON state. Thus, quite generally, the mean abundance of mRNA from gene i in the population of many cells, measured as transcripts per million (TPM) in our RNA-seq, can be expressed as $\mathrm{TPM}_i \propto J_i f_i / d_i$ (Equation S3). Therefore, TPMs averaged over

many genes with uncorrelated J_i, f_i, d_i should be proportional to f, the average fraction of these genes in the ON state. Figure 5B shows that this is indeed true for 11x OE and KO when we take ATAC-seq-derived promoter accessibility *a* as a proxy for f. Based on the paradigm model, MeCP2 could in principle regulate transcription by affecting initiation, elongation or the fraction of cells with specific genes in the ON state (Figure 5A).

In our first model – the ON/OFF Condensation Model – MeCP2 affects the proportion f of genes in the ON state by condensing chromatin, for example by recruiting remodelling enzymes (Figure 5C). In accordance with our paradigm model we assume that f is proportional to the fraction of "open" (accessible) promoters as measured by accessibility a. Figure 5D shows that a decreases with increasing promoter mCG density, perhaps reflecting the known inhibitory effect of promoter methylation on transcription. Notably, accessibility of gene bodies also declines with increasing MeCP2 concentration but only a weak dependence on gene body methylation is discernible (Figure S5A). A possible interpretation of these results is that promoter accessibility (and thus the fraction f_i of cells with gene i in the active state) depends on both the level of MeCP2 and the level of gene methylation. However, Log2FC of the expression ratio OE/KO predicted by promoter accessibility from the model (Figure S5B; black line) does not align with the Log2FC curve from the RNA-seq data (Figure S5B; red line). A similar plot focussed on the mCG density in gene bodies shows a clear Log2FC gene expression decay with increasing mCG density in the RNA-seq data, whereas the curve predicted from the model is almost flat (Figure 5E). We therefore conclude that the change in the proportion of ON/OFF genes caused by chromatin condensation cannot satisfactorily explain most of the observed correlation between expression and mCG density in gene bodies.

A second model – the Initiation Model – proposes that MeCP2 bound downstream of the promoter affects transcription initiation via mechanism such as interfering with binding of Pol II and cofactors or Pol II pausing (Figure 5C) (Jonkers and Lis, 2015). In this model, we expect Log2FC to correlate better with mCG density near the promoter region than far away from the promoter. This does not agree with Figure 4 which shows the opposite effect. To confirm that this is not an artefact of differences in the accuracy with which mCG density can be determined in short promoter regions versus much longer gene bodies, we selected 2kbp-long regions around the TSS and compared them with same-length regions 40kb away (TSS+40kb), for ~3000 genes longer than 50kb. A potential confounding factor would be long-range correlations in mCG density which might cause Log2FC to correlate with both the promoter and gene body methylation even in the absence of a causal relationship. This correlation is very weak, however ($r = 0.08$, Figure S5C). Plotting Log2FC(OE 11x/KO) versus TSS+40±1kb (gene body) mCG density, showed the characteristic decrease in gene expression with increasing gene body methylation (Figure S5D). In contrast, the dependence on promoter methylation (TSS±1kb) was weak, offering little support for the Initiation Model. The non-zero slope in OE 11x may indicate a weak effect of MeCP2 on promoter activity, or it may reflect residual correlations between promoter and gene body methylation.

Since our analysis points to MeCP2 affecting transcription predominantly via gene body methylation, a reasonable expectation is that MeCP2 affects some aspect of the elongation phase of transcription. In our third model – the Detachment Model (Figure 5F) – MeCP2 causes Pol II to abort transcription either by directly blocking further polymerase progress or

by indirectly causing the polymerase to detach from DNA, e.g., via chromatin modification. Log2FC in this model (Methods) is a function of the total number of methylated CGs (N_{mCG}) in the gene: $\mathrm{Log2FC} = \exp(-\gamma M N_{mCG})$, where M is MeCP2 concentration relative to WT, and the parameter γ is proportional to the probability that Pol II aborts transcription when it collides with MeCP2. The dependence on the total number of mCGs in the gene is due to each "roadblock" increasing the probability of premature termination. We calculated the model's only parameter γ by fitting the model to the Log2FC (KO/WT) data (Figure 5G, left panel). We found that the model failed to reproduce the Log2FC vs N_{mCG} relationship for the OE 11X cell line (Figure 5G, right panel). The model also incorrectly predicts the observed relationship between Log2FC and mCG density in gene bodies (Figures S5E,F). We conclude that the Detachment Model can be rejected.

MeCP2 creates "slow sites" that impede transcriptional elongation

Finally, we considered a hypothesis that Pol II slows down when it encounters an obstruction in the form of MeCP2 or a reversible chemical modification of chromatin induced by MeCP2. This leads to Pol II queuing in front of the obstruction (Figure 6A), the density of which is equal to the density of MeCP2 on the DNA. Mathematical modelling (Methods) shows that, regardless of model details, the maximum slope of the Log2FC curves from Figure 4C should be proportional to the concentration of MeCP2, in agreement with the experimental data (Figure 4D). In the Slow Sites Congestion Model (our fourth model), which is based on exclusion processes (Blythe and Evans, 2007; Derrida, 1998; Sahoo et al., 2014; Turci et al., 2013), MeCP2 creates "slow sites" at methylated CGs. The parameters are: the fraction p of mCGs bound by MeCP2 (estimated from ATAC-seq), the speed v of Pol II far from affected mCGs, the speed v_s at affected mCGs, and (specific to each gene) the length L of the gene, the density ρ of methylated CGs, and the transcription initiation rate α. We took $v = 60$bp/s, close to the maximum observed speed of Pol II in human cells (Fuchs et al., 2014; Veloso et al., 2014). In the biologically realistic limit $k_{ON}, k_{OFF} \ll \alpha$, the transcription rate J predicted by the model is determined by three parameters: $\alpha, v_s/v$, and $p\rho$. Figure 6B shows the transcription rate for OE 11x predicted by the model as a function of initiation rate α, for different ρ and for $v_s/v = 0.05$. The assumed value of $v_s/v = 0.05$ anticipates a temporary MeCP2-dependent "roadblock" lasting on average 20s, which is compatible with the reported *in vivo* residence time of MeCP2 on chromatin (25-40s (Klose et al., 2005)). All curves in Figure 6B have a characteristic shape: a linear relationship $J \approx \alpha$ for small α, followed by saturation at high initiation rates. Saturation occurs due to congestion: polymerases form queues in front of slow sites (SI Videos 1,2). A similar behaviour occurs in a fifth model (a modification of the Slow Sites Congestion Model) – the Dynamical Obstacles Congestion Model – in which MeCP2 molecules bind/unbind to mCG and directly block Pol II (Figure S6E,F, SI Video 3).

To estimate the rates α we fitted both models (fourth and fifth) to Log2FC(OE 11x/OE ctr) versus gene-body mCG density (Figures 6C and S6G). The predictions of both models agree well with the experimental data for OE 4x and KO (Figures 6C and S6G). The slopes of the Log2FC plots from the models also agree with the experimental slope for all seven cell lines (Figures 6D and S6H).

In summary, of the five models we considered, only the two Congestion Models (Slow Sites and Dynamical Obstacles) cannot be rejected based on the available data, and we consider the slowing down of Polymerase II to be a plausible mechanism by which MeCP2 affects transcription (Figure 6E).

MeCP2 binding to both DNA and NCoR are essential to slow down RNA Pol II

Under the Congestion Models, MeCP2 may impede RNA polymerase II progression directly by simple steric interference. Alternatively, it may induce an altered local chromatin structure that retards RNA polymerisation indirectly, for example by increasing the affinity of nucleosomes for DNA by deacetylation (Zentner and Henikoff, 2013). Aware that MeCP2 overexpression caused robust effects on transcription, we overexpressed mutant forms of MeCP2 (R111G and R306C) to a similar extent in order to distinguish these possibilities (Figures S7A-C). The Rett mutant R111G is unable to bind methylated DNA (Kudo et al., 2003), while Rett mutant R306C has lost the ability to recruit the histone deacetylase complex NCoR (Kruusvee et al., 2017; Lyst et al., 2013) (Figure 7A). The R111G mutant protein was 7-fold over-expressed and R306C was 11-fold over-expressed compared to WT control (Figures 7B and S7D,E). In the case of R111G, overexpression is not expected to affect transcription, regardless of whether MeCP2 interferes directly or indirectly with transcription elongation, as both effects would require binding to DNA. Accordingly, we observed no mCG-density dependent transcriptional changes due to overexpression of MeCP2-R111G (Figure 7C). The R306C mutant, on the other hand, should repress transcription if inhibition is directly due to MeCP2 binding to DNA. Again, we observed that 11-fold overexpression of MeCP2-R306C relative to WT MeCP2 gave a predominantly flat curve, indicating a significant loss of DNA methylation-dependent repression (Figure 7D). The presence of a weak slope suggests that minimal DNA methylation dependent repression persists, perhaps due to direct interference of DNA-bound MeCP2-R306C with transcription. The predominant conclusion is that MeCP2 binding to methylated cytosines is essential for slowing down elongation, with recruitment of the NCoR co-repressor complex playing an important role, as R306C, like R111G, does not fall on the line defining the linear relationship between gene repression and MeCP2 concentration (Figure 7E). On balance these findings favour an indirect mechanism of repression, whereby MeCP2 alters the chromatin state to impede the transcription process.

DISCUSSION

Accumulating evidence indicates that MeCP2 influences the expression of large numbers of genes in the brain. While the advent of highly replicated, well-controlled RNA-seq analyses has allowed the detection of the relationship between MeCP2 binding and restraint of transcription (Gabel et al., 2015; Kinde et al., 2016; Lagger et al., 2017), the small magnitude of differences in transcription between wildtype and mutant cells or tissues makes it difficult to distinguish the underlying mechanisms. To overcome this limitation, we have intimately combined experimental design and data capture with mathematical modelling to derive quantitative predictions with the aim of distinguishing competing hypotheses. To generate a consistent and relevant source of experimental data, we modified a genetically homogeneous line of cultured human neuronal progenitor cells to generate neuronal lines expressing widely disparate levels of MeCP2. Despite this difference in MeCP2 concentrations, each line retained the ability to rapidly differentiate into neurons.

Combined molecular characterisation and mathematical modelling revealed new features of MeCP2 binding to chromatin. Notably, we used ATAC-seq to visualise for the first time an *in vivo* footprint of MeCP2 binding to genomic mCG and mCA sites. Footprint intensity increased with MeCP2 concentration, providing direct evidence that the primary mode of binding to genomic DNA *in vivo* is dependent on methylation of cytosine in these short sequences. Mathematical modelling of quantitative ChIP-seq and ATAC-seq data provided independent estimates of the fraction of all potential binding sites that are occupied by MeCP2. Both methodologies agree that *in vivo* only ~6% of all mCGs are occupied in OE 11x neurons, which is far from theoretical saturation. A potential contributory factor could be restricted accessibility to DNA due to nucleosomes, which may allow MeCP2 to associate exclusively with exposed inter-nucleosomal linker DNA. It was initially suggested that MeCP2 can bind to DNA on the surface of a nucleosome (Chandler et al., 1999), but more recent studies have emphasised its association with linkers and exclusion from nucleosomal DNA (Ishibashi et al., 2008; Nikitina et al., 2007; Thambirajah et al., 2011). Our results support the latter view.

Having developed a data-led mathematical model of MeCP2 binding to chromatin, we used this information to explore its effects on transcription in neurons expressing various levels of MeCP2, ranging from zero to 11x the WT amount. A strength of our approach is that we included all the transcription data, rather than focussing only on genes whose change in transcription was arbitrarily classified as significant. In general, increasing MeCP2 abundance caused progressively larger effects on gene expression, allowing us to clearly relate the degree of transcriptional disturbance to the density of mCG in gene bodies. This high-level relationship extends previous evidence derived from mouse brain (Kinde et al., 2016; Lagger et al., 2017), but does not shed light on the mechanisms at work. To address this issue, we evaluated RNA-seq data against a series of mechanistic models. Models whereby MeCP2 reduces initiation of transcription were not favoured, suggesting that elongation of RNA Pol II through the transcription unit may be affected. We rejected the possibility of premature transcription termination, and finally considered a model in which MeCP2 slows down the progress of RNA Pol II without displacing it. This last model reproduced the data very well.

Finally, we asked whether the inhibition of transcript elongation could be direct, for example due to interference with RNA Pol II progression by bound molecules of MeCP2, or if its effects are indirectly mediated by other mechanisms. To address this issue, we over-expressed Rett syndrome-causing mutant forms of MeCP2 that lacked either DNA binding capacity or the ability to interact with the NCoR co-repressor complexes. Under either model, loss of the interaction between methylated DNA and MeCP2 would abolish the relationship between MeCP2 binding sites and transcription, a scenario we indeed observed. Importantly, a missense mutation that greatly reduces the MeCP2-NCoR interaction also violated the established relationship between MeCP2 concentration and transcriptional inhibition. This result is compatible with the view that MeCP2 binding affects the chromatin template indirectly so as to impede transcription. A candidate mediator of this effect is histone modification, in particular histone acetylation. The NCoR1 and NCoR2 complexes both recruit histone deacetylases whose action would increase the affinity between DNA and nucleosomes and thus inhibit transcription (Zentner and Henikoff, 2013). In fact, cell transfection assays using MeCP2 and methylated reporters have demonstrated that repression depends upon histone deacetylase activity (Nan et al., 1997; 1998).

In summary, a close alliance between mathematical modelling and quantitative molecular characterisation of human neurons has allowed us to discriminate molecular mechanisms underlying the effects of MeCP2 on gene expression. We propose that our approach will illuminate the mechanisms by which other global regulators choreograph gene expression programmes.

Acknowledgments

We thank Tanja Waldmann for introducing us to the LUHMES cell line, Beatrice Alexander-Howden for technical support, David Kelly for microscope assistance, Martin Waterfall for help with FACS and Jim Selfridge for help with preparing samples for HPLC. We thank Sabine Lagger and John Connelly for critical assessment of the manuscript. The work has made use of resources provided by the Edinburgh Compute and Data Facility (ECDF; www.ecdf.ed.ac.uk) and was supported by a Wellcome Trust Programme Grant and Investigator Award to AB. JCW was supported by a grant from the Rett Syndrome Research Trust. RS was supported by a Wellcome Trust 4 year PhD studentship. BW was supported by a Personal Research Fellowship from the Royal Society of Edinburgh.

Author contributions

Conceptualization, A.B., J.C.-W., R.S., B.W.; Methodology, J.C.-W., P.G., O.P., R.S., B.W.; Software, K.C., P.G., B.W., S.W.; Formal analysis, K.C., J.C.-W., B.R., B.W., S.W.; Investigation, K.C., J.C.-W., O.P., B.R., R.S., B.W., S.W., M.Y.; Visualization, K.C., J.C.-W., R.S., B.W., S.W.; Writing – Original draft, A.B., J.C.-W., R.S., B.W.; Writing – Review & Editing: all authors; Supervision, A.B. and B.W.; Funding acquisition, A.B.

References

Amir, R.E., Van den Veyver, I.B., Wan, M., Tran, C.Q., Francke, U., and Zoghbi, H.Y. (1999). Rett syndrome is caused by mutations in X-linked MECP2, encoding methyl-CpG-binding protein 2. Nat Genet *23*, 185–188.

Baubec, T., Ivánek, R., Lienert, F., and Schübeler, D. (2013). Methylation-Dependent and -Independent Genomic Targeting Principles of the MBD Protein Family. Cell *153*, 480–492.

Berry, S., Dean, C., and Howard, M. (2017). Slow Chromatin Dynamics Allow Polycomb Target Genes to Filter Fluctuations in Transcription Factor Activity. Cell Systems *4*, 445–457.e448.

Bird, A., Nan, X., Ng, H.-H., Johnson, C.A., Laherty, C.D., Turner, B.M., and Eisenman, R.N. (1998). Transcriptional repression by the methyl-CpG-binding protein MeCP2 involves a histone deacetylase complex. Nature *393*, 386–389.

Blythe, R.A., and Evans, M.R. (2007). Nonequilibrium steady states of matrix-product form: a solver's guide. J. Phys. a: Math. Theor. *40*, R333–R441.

Bolger, A.M., Lohse, M., and Usadel, B. (2014). Trimmomatic: a flexible trimmer for Illumina sequence data. Bioinformatics *30*, 2114–2120.

Buenrostro, J.D., Giresi, P.G., Zaba, L.C., Chang, H.Y., and Greenleaf, W.J. (2013). Transposition of native chromatin for fast and sensitive epigenomic profiling of open chromatin, DNA-binding proteins and nucleosome position. Nat Meth *10*, 1213–1218.

Buenrostro, J.D., Wu, B., Chang, H.Y., and Greenleaf, W.J. (2015). ATAC-seq: A Method for Assaying Chromatin Accessibility Genome-Wide (Hoboken, NJ, USA: John Wiley & Sons, Inc.).

Chandler, S.P., Guschin, D., Landsberger, N., and Wolffe, A.P. (1999). The Methyl-CpG Binding Transcriptional Repressor MeCP2 Stably Associates with Nucleosomal DNA. Biochemistry *38*, 7008–7018.

Chen, R.Z., Akbarian, S., Tudor, M., and Jaenisch, R. (2001). Deficiency of methyl-CpG binding protein-2 in CNS neurons results in a Rett-like phenotype in mice. Nat Genet *27*, 327–331.

Derrida, B. (1998). An exactly soluble non-equilibrium system: The asymmetric simple exclusion process. Physics Reports *301*, 65–83.

Dobin, A., Davis, C.A., Schlesinger, F., Drenkow, J., Zaleski, C., Jha, S., Batut, P., Chaisson, M., and Gingeras, T.R. (2013). STAR: ultrafast universal RNA-seq aligner. Bioinformatics *29*, 15–21.

Epshtein, V. (2003). Transcription through the roadblocks: the role of RNA polymerase cooperation. Embo J. *22*, 4719–4727.

Fuchs, G., Voichek, Y., Benjamin, S., Gilad, S., Amit, I., and Oren, M. (2014). 4sUDRB-seq: measuring genomewide transcriptional elongation rates and initiation frequencies within cells. Genome Biol *15*, R69.

Gabel, H.W., Kinde, B., Stroud, H., Gilbert, C.S., Harmin, D.A., Kastan, N.R., Hemberg, M., Ebert, D.H., and Greenberg, M.E. (2015). Disruption of DNA-methylation-dependent long gene repression in Rett syndrome. Nature 1–21.

Guo, J.U., Su, Y., Shin, J.H., Shin, J., Li, H., Bin Xie, Zhong, C., Hu, S., Le, T., Fan, G., et al. (2014). Distribution, recognition and regulation of non-CpG methylation in the adult mammalian brain. Nat. Neurosci. *17*, 215–222.

Guy, J., Hendrich, B., Holmes, M., Martin, J.E., and Bird, A. (2001). A mouse Mecp2-null mutation causes neurological symptoms that mimic Rett syndrome. Nat Genet *27*, 322–326.

Hao, N., Palmer, A.C., Dodd, I.B., and Shearwin, K.E. (2017). Directing traffic on DNA—How transcription factors relieve or induce transcriptional interference. Transcription *8*, 120–125.

Hendrich, B., and Bird, A. (1998). Identification and characterization of a family of mammalian methyl-CpG binding proteins. Molecular and Cellular Biology *18*, 6538–6547.

Ho, K.L., McNae, I.W., Schmiedeberg, L., Klose, R.J., Bird, A.P., and Walkinshaw, M.D. (2008). MeCP2 Binding to DNA Depends upon Hydration at Methyl-CpG. Molecular Cell *29*, 525–531.

Jonkers, I., and Lis, J.T. (2015). Getting up to speed with transcription elongation by RNA polymerase II. Nat Rev Mol Cell Biol *16*, 167–177.

Kinde, B., Wu, D.Y., Greenberg, M.E., and Gabel, H.W. (2016). DNA methylation in the gene body influences MeCP2-mediated gene repression. Proc. Natl. Acad. Sci. U.S.A. *113*, 15114–15119.

Kishi, N., and Macklis, J.D. (2004). MECP2 is progressively expressed in post-migratory neurons and is involved in neuronal maturation rather than cell fate decisions. Mol. Cell. Neurosci. *27*, 306–321.

Klose, R.J., Sarraf, S.A., Schmiedeberg, L., McDermott, S.M., Stancheva, I., and Bird, A.P. (2005). DNA Binding Selectivity of MeCP2 Due to a Requirement for A/T Sequences Adjacent to Methyl-CpG. Molecular Cell *19*, 667–678.

Krueger, F., and Andrews, S.R. (2011). Bismark: a flexible aligner and methylation caller for Bisulfite-Seq applications. Bioinformatics *27*, 1571–1572.

Krug, J. (1991). Boundary-induced phase transitions in driven diffusive systems. Phys. Rev. Lett. *67*, 1882–1885.

Kruusvee, V., Lyst, M.J., Taylor, C., Tarnauskaitė, Ž., Bird, A.P., and Cook, A.G. (2017). Structure of the MeCP2–TBLR1 complex reveals a molecular basis for Rett syndrome and related disorders. Proc. Natl. Acad. Sci. U.S.A. *114*, E3243–E3250.

Kudo, S., Nomura, Y., Segawa, M., Fujita, N., Nakao, M., Schanen, C., and Tamura, M. (2003). Heterogeneity in residual function of MeCP2 carrying missense mutations in the methyl CpG binding domain. J Med Genet *40*, 487–493.

Lagger, S., Connelly, J.C., Schweikert, G., Webb, S., Selfridge, J., Ramsahoye, B.H., Yu, M., He, C., Sanguinetti, G., Sowers, L.C., et al. (2017). MeCP2 recognizes cytosine methylated tri-nucleotide and di-nucleotide sequences to tune transcription in the mammalian brain. PLoS Genet. *13*, e1006793.

Lewis, J.D., Meehan, R.R., Henzel, W.J., Maurer-Fogy, I., Jeppesen, P., Klein, F., and Bird, A. (1992a). Purification, sequence, and cellular localization of a novel chromosomal protein that binds to methylated DNA. Cell *69*, 905–914.

Lewis, J.D., Meehan, R.R., Henzel, W.J., Maurer-Fogy, I., Jeppesen, P., Klein, F., and Bird, A. (1992b). Purification, sequence, and cellular localization of a novel chromosomal protein that binds to Methylated DNA. Cell *69*, 905–914.

Li, B., Carey, M., and Workman, J.L. (2007). The Role of Chromatin during Transcription. Cell *128*, 707–719.

Li, H. (2013). Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. Genomics *1303*.

Li, Y., Wang, H., Muffat, J., Cheng, A.W., Orlando, D.A., Lovén, J., Kwok, S.-M., Feldman, D.A., Bateup, H.S., Gao, Q., et al. (2013). Global Transcriptional and Translational Repression

in Human-Embryonic-Stem-Cell-Derived Rett Syndrome Neurons. Cell Stem Cell *13*, 446–458.

Liang, K., and Keles, S. (2012). Normalization of ChIP-seq data with control. BMC Bioinformatics *13*, 199.

Liao, Y., Smyth, G.K., and Shi, W. (2014). featureCounts: an efficient general purpose program for assigning sequence reads to genomic features. Bioinformatics *30*, 923–930.

Lin, C.Y., Lovén, J., Rahl, P.B., Paranal, R.M., Burge, C.B., Bradner, J.E., Lee, T.I., and Young, R.A. (2012). Transcriptional Amplification in Tumor Cells with Elevated c-Myc. Cell *151*, 56–67.

Lister, R., Mukamel, E.A., Nery, J.R., Urich, M., Puddifoot, C.A., Johnson, N.D., Lucero, J., Huang, Y., Dwork, A.J., Schultz, M.D., et al. (2013a). Global Epigenomic Reconfiguration During Mammalian Brain Development. Science *341*, 1237905–1237905.

Lister, R., Mukamel, E.A., Nery, J.R., Urich, M., Puddifoot, C.A., Johnson, N.D., Lucero, J., Huang, Y., Dwork, A.J., Schultz, M.D., et al. (2013b). Global Epigenomic Reconfiguration During Mammalian Brain Development. Science *341*, 1237905–1237905.

Love, M.I., Huber, W., and Anders, S. (2014). Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. Genome Biol *15*, 31.

Lovelace, D.L., Gao, Z., Mutoji, K., Song, Y.C., Ruan, J., and Hermann, B.P. (2016). The regulatory repertoire of PLZF and SALL4 in undifferentiated spermatogonia. Development *143*, 1893–1906.

Lyst, M.J., Ekiert, R., Ebert, D.H., Merusi, C., Nowak, J., Selfridge, J., Guy, J., Kastan, N.R., Robinson, N.D., de Lima Alves, F., et al. (2013). Rett syndrome mutations abolish the interaction of MeCP2 with the NCoR/SMRT co-repressor. Nat. Neurosci. *16*, 898–902.

Nan, X., Campoy, F.J., and Bird, A. (1997). MeCP2 is a transcriptional repressor with abundant binding sites in genomic chromatin. Cell *88*, 471–481.

Nan, X., Meehan, R.R., and Bird, A. (1993). Dissection of the methyl-CpG binding domain from the chromosomal protein MeCP2. Nucleic Acids Research *21*, 4886–4892.

Nan, X., Tate, P., Li, E., and Bird, A. (1996). DNA methylation specifies chromosomal localization of MeCP2. Molecular and Cellular Biology *16*, 414–421.

Nan, X., Ng, H.-H., Johnson, C.A., Laherty, C.D., Turner, B.M., Eisenman, R.N., and Bird, A. (1998). Transcriptional repression by the methyl-CpG-binding protein MeCP2 involves a histone deacetylase complex. Nature *393*, 386–389.

Ouararhni, K., Hadj-Slimane, R., Ait-Si-Ali, S., Robin, P., Mietton, F., Harel-Bellan, A., Dimitrov, S., and Hamiche, A. (2006). The histone variant mH2A1.1 interferes with transcription by down-regulating PARP-1 enzymatic activity. Genes Dev. *20*, 3324–3336.

Patro, R., Mount, S.M., and Kingsford, C. (2014). Sailfish enables alignment-free isoform quantification from RNA-seq reads using lightweight algorithms. Nat Biotechnol *32*, 462–464.

Perrenoud, A., and Sauer, U. (2005). Impact of global transcriptional regulation by ArcA, ArcB, Cra, Crp, Cya, Fnr, and Mlc on glucose catabolism in Escherichia coli. Journal of Bacteriology *187*, 3171–3179.

Rabani, M., Levin, J.Z., Fan, L., Adiconis, X., Raychowdhury, R., Garber, M., Gnirke, A., Nusbaum, C., Hacohen, N., Friedman, N., et al. (2011). Metabolic labeling of RNA uncovers principles of RNA production and degradation dynamics in mammalian cells. Nat Biotechnol *29*, 436–442.

Ramírez, F., Ryan, D.P., Grüning, B., Bhardwaj, V., Kilpert, F., Richter, A.S., Heyne, S., Dündar, F., and Manke, T. (2016). deepTools2: a next generation web server for deep-sequencing data analysis. Nucleic Acids Research *44*, W160–W165.

Ramocki, M.B., Tavyev, Y.J., and Peters, S.U. (2010). The MECP2 Duplication Syndrome. American Journal of Medical Genetics. Part A *152A*, 1079–1088.

Sahoo, M., Dong, J., and Klumpp, S. (2014). Dynamic blockage in an exclusion process. J. Phys. a: Math. Theor. *48*, 015007.

Schep, A.N., Buenrostro, J.D., Denny, S.K., Schwartz, K., Sherlock, G., and Greenleaf, W.J. (2015). Structured nucleosome fingerprints enable high-resolution mapping of chromatin architecture within regulatory regions. Genome Res. *25*, 1757–1770.

Scholz, D., Pöltl, D., Genewsky, A., Weng, M., Waldmann, T., Schildknecht, S., and Leist, M. (2011). Rapid, complete and large-scale generation of post-mitotic neurons from the human LUHMES cell line. J Neurochem *119*, 957–971.

Shah, R.R., Cholewa-Waclaw, J., Davies, F.C.J., Paton, K.M., Chaligne, R., Heard, E., Abbott, C.M., and Bird, A.P. (2016). Efficient and versatile CRISPR engineering of human neurons in culture to model neurological disorders. Wellcome Open Res *1*, 13.

Shahbazian, M.D. (2002). Insight into Rett syndrome: MeCP2 levels display tissue- and cell-specific differences and correlate with neuronal maturation. Hum. Mol. Genet. *11*, 115–124.

Shahrezaei, V., Ollivier, J.F., and Swain, P.S. (2008). Colored extrinsic fluctuations and stochastic gene expression. Molecular Systems Biology *4*, 196.

Skene, P.J., Illingworth, R.S., Webb, S., Kerr, A.R.W., James, K.D., Turner, D.J., Andrews, R., and Bird, A.P. (2010). Neuronal MeCP2 is expressed at near histone-octamer levels and globally alters the chromatin state. Molecular Cell *37*, 457–468.

Sneppen, K., Dodd, I.B., Shearwin, K.E., Palmer, A.C., Schubert, R.A., Callen, B.P., and Egan, J.B. (2005). A Mathematical Model for Transcriptional Interference by RNA Polymerase Traffic in Escherichia coli. Journal of Molecular Biology *346*, 399–409.

Sperlazza, M.J., Bilinovich, S.M., Sinanan, L.M., Javier, F.R., and Williams, D.C., Jr (2017). Structural Basis of MeCP2 Distribution on Non-CpG Methylated and Hydroxymethylated DNA. Journal of Molecular Biology *429*, 1581–1594.

Turci, F., Parmeggiani, A., Pitard, E., Romano, M.C., and Ciandrini, L. (2013). Transport on a lattice with dynamical defects. Phys. Rev. E *87*, 012705.

Veloso, A., Kirkconnell, K.S., Magnuson, B., Biewen, B., Paulsen, M.T., Wilson, T.E., and Ljungman, M. (2014). Rate of elongation by RNA polymerase II is associated with specific gene features and epigenetic modifications. Genome Res. *24*, 896–905.

Woo, H., Dam Ha, S., Lee, S.B., Buratowski, S., and Kim, T. (2017). Modulation of gene expression dynamics by co-transcriptional histone methylations. Exp Mol Med *49*, e326–e326.

Workman, J.L., and Kingston, R.E. (1998). Alteration of nucleosome structure as a mechanism of transcriptional regulation. Annu. Rev. Biochem. *67*, 545–579.

Xu, H., Handoko, L., Wei, X., Ye, C., Sheng, J., Wei, C.-L., Lin, F., and Sung, W.-K. (2010). A signal-noise model for significance analysis of ChIP-seq with negative control. Bioinformatics *26*, 1199–1204.

Yazdani, M., Deogracias, R., Guy, J., Poot, R.A., Bird, A., and Barde, Y.-A. (2012). Disease Modeling Using Embryonic Stem Cells: MeCP2 Regulates Nuclear Size and RNA Synthesis in Neurons. Stem Cells 30(10), 2128-2139.

Yu, M., Hon, G.C., Szulwach, K.E., Song, C.-X., Zhang, L., Kim, A., Li, X., Dai, Q., Shen, Y., Park, B., et al. (2012). Base-resolution analysis of 5-hydroxymethylcytosine in the Mammalian genome. Cell *149*, 1368–1380.

Zentner, G.E., and Henikoff, S. (2013). Regulation of nucleosome dynamics by histone modifications. Nat Struct Mol Biol *20*, 259–266.

METHODS

Experimental model and subject details

We used LUHMES-derived human neurons as a convenient cellular system characterized by rapid and effective differentiation (~5 days) and maturation times (~10 days) to an electrically active state (Figure S1A and (Scholz et al., 2011)). LUHMES cells are a sub-clone of the MESC.1 cell line, isolated from mesencephalon of an 8-week old embryo and immortalized with v-myc under the control of a TET-OFF system to maintain their proliferating capacity (Scholz et al., 2011). As immortalized neuronal precursors (day 0), they express Sox2 and are negative for mature neuronal markers such as MAP2 and TH (Figure S1B). After the addition of tetracycline in the culture medium to turn off *v-myc* expression, LUHMES cells differentiate and express neuronal markers such MAP2, NeuN (Fox-3) and dopaminergic markers such as TH, DRD2, and DAT (Figures S1B,C). MeCP2 is progressively upregulated during the course of differentiation and its maximal level is achieved around day 9 (Figure S1D).

MeCP2 level manipulation

To create two independent *MECP2* knock-out lines, we used CRISPR-mediated gene disruption (Shah et al., 2016). Briefly, cells were nucleofected with two plasmids; one containing Cas9 and a guide RNA targeting exon 3 of *MECP2* gene, and another encoding Puromycin resistance (Table S2, Figure S1F).

For generation of MeCP2 knock-downs, several shRNAs against MeCP2 were designed using Sigma-Aldrich Mission shRNA online software. Two shRNAs were chosen and cloned into pLKO.1 vector including scrambled shRNA as a control and lentiviruses were created (Table S2, Figure S1G). Knock-down cell lines were analysed as pool of cells.

To increase the level of MeCP2 we created lentiviruses expressing MeCP2 from two alternative promoters in the pLKO.1 vector: Synapsin and cytomegalovirus (CMV)(Figure S1H). The human Synapsin promoter drove expression of the MeCP2 E2 isoform, while the CMV promoter drove the MeCP2 E1 isoform. As a control, we expressed GFP under the control of the CMV promoter (Figures S1I). The Synapsin promoter becomes active on day 5 of LUHMES differentiation (Figure S1M - middle panels; see insets), while the CMV promoter is expressed from day 0 which allowed us to generate cell lines expressing different levels of MeCP2 overexpression, as follows. Cells overexpressing MeCP2 at 3-fold wildtype levels were obtained by infecting cells with lentiviral particles containing the Synapsin-MeCP2 construct. The resulting cells (OE 3x) were analysed as a pool. OE 4x and OE 11x was generated by infecting cells with lentiviruses containing the Synapsin-MeCP2 and CMV-MeCP2 constructs, respectively (Figure S1M – upper and middle panels; see insets for GFP expression). OE control cells were made by infecting cells with viruses containing the CMV-GFP construct (S1M – bottom panels; see insets). All three cell lines (OE 4x, OE 11x and OE ctr) were clonally selected by FACS into 96-well plates. OE R111G and R306C lines were generated by infecting cells with lentiviruses containing the appropriate CMV-mutated MeCP2 construct (Figure S7A) and selected as previously described. The mutations were introduced into plasmid using QuickChange II site-directed mutagenesis kit (Agilent Technologies; 200523).

Method Details

LUHMES culture and differentiation

The procedure for culture and differentiation of the LUHMES cell line was previously described (Scholz, Poltl et al. 2011). Briefly, proliferating LUHMES cells were cultured on Nunc plasticware coated with 40 µg/ml Poly-L-ornithine (Sigma-Aldrich; P-3655-100mg) and 1µg/ml Fibronectin (Sigma-Aldrich; F-1141-5mg) in medium for proliferating cells: Advanced DMEM (Life Technologies; 12634-010) plus N2 supplement (Life Technologies; 17502-048), 2mM L-Glutamine (Sigma-Aldrich; G7513), and 40 µg/ml bFGF (R&D Systems; 4114-TC-1 mg). For differentiation, LUHMES cells were seeded on 40 µg/ml Poly-L-Ornithine and 1 µg/ml Fibronectin coated plastic-ware at a density of 3-4 x 10^4 cells/cm^2 in medium for proliferating cells. Next day after seeding, the medium was changed to the differentiation medium containing Advanced DMEM, 2mM L-Glutamine, N2 supplement, 1mM cAMP (Sigma-Aldrich; D0627), 1µg/ml Tetracycline (Sigma-Aldrich; T-7660) and 2 ng/ml GDNF (R&D Systems; 212-

GD-50 µg). Two days later, cells were trypsinised and seeded at a density of $1.1x10^5$ cells/cm^2 in the differentiation medium for the final differentiation.

Lentiviral particle preparation and LUHMES infection

Lentiviruses were produced in HEK293FT cells (Life Technologies; R700-07) by retrograde transfection of vectors: 7.5 µg pLKO.1; 4.6 µg psPax2; 2.8 µg pMD2G using Lipofectamine 2000 (Life Technologies; 11668027). Briefly, for one transfection, plasmids in appropriate ratios were added to 1.5 ml OptiMEM (Life Technologies; 31985062) and 36 µl of Lipofectamine 2000 was mixed with 1.5 ml of OptiMEM. Both solutions were combined and incubated for 20 min at room temperature. Meanwhile, HEK293FT cells were trypsinised and concentration-adjusted to $1.2x10^6$/ml. 5 ml of cell suspension was added to a 10 cm dish; the transfection mix was then added and the culture was left overnight. Cells were left for 72 hours, after which the supernatant was collected, and viral particles were precipitated using a PEG Virus Precipitation kit (BioVision) following the manufacturer's protocol. Viruses were aliquoted and either used directly or frozen at -80ºC. LUHMES cells were infected with lentiviral particles overnight. The next day, cells were selected with 0.5 µg/ml Puromycin (Life Technologies; A11138-03), and concentration of antibiotic was reduced the following day to 0.25 µg/ml. The population of resistant cells was either analysed as a pool or FACS-sorted into a 96-well plate in order to isolate single clones.

Western blot, qPCR, immunofluorescence, live cell imaging

For protein analysis cells were either lysed to give a whole cell extract or nuclei were isolated. For whole cell extracts cells were homogenized in the NE1 buffer (20 mM HEPES pH 7.9, 10 mM KCl, 1 mM $MgCl_2$, 0.5 mM DTT, 0.1% Triton X, 20% glycerol, 2mM PMSF, protease inhibitor) and treated with Benzonase (Sigma) for 15 min at RT to remove DNA. Protein concentration was measured (Bradford, Bio-Rad) and loading buffer (5x, 250 mM Tris·HCl pH 6.8, 10% SDS, 30% [v/v] Glycerol, 10 mM DTT, 0.05% [w/v] Bromophenol Blue) was added. Samples were boiled for 5 min prior to loading 40 µg of protein extract onto a polyacrylamide gel.

To prepare nuclei, cells were lysed in hypotonic buffer (0.32 M Sucrose, 10 mM Tris-HCl pH 8, 3 mM $CaCl_2$, 2 mM MgOAc, 0.1 mM EDTA, 0.5% NP-40). Nuclei were washed with NP-40-free buffer, counted, centrifuged, and resuspended in an appropriate volume of 20% glycerol in PBS. Nuclei were treated with Benzonase, loading buffer was added and samples were boiled for 5 min. Nuclear lysates ($1.6x10^5$ per lane) were run on the gradient 4-15% SDS-PAGE gels (Mini-Protean, Bio-Rad) in running buffer (25 mM Tris, 190 mM glycine, 0.1% SDS) at 200 V for ~40 min alongside a protein size marker (PageRuler, Thermo Scientific). Proteins were transferred onto a Nitrocellulose membrane using wet transfer method for 1 hour at 200V. The membrane was stained first with Ponceau S solution for quality check and then treated with blocking buffer (PBS, 1% PVP, 1% non-fat dried milk, 0.1% Tween 20, 0.01% NaN_3) for 30 min at RT. The membrane was incubated for 1 hour with primary antibodies (see Table S3) at room temperature and then 1 hour with secondary antibodies conjugated with either IRDye 700DX or IRDye 800CW (LI-COR). The membrane was washed with PBS containing 0.1% Tween 20 after adding each antibody and was imaged using Odyssey CLx (LI-COR).

For qPCR analysis, total RNA was isolated using RNeasy kit (Qiagen) and treated with DNase I (DNA free kit, Ambion) to remove genomic DNA. The efficient removal of genomic DNA in the RNA samples was tested by performing PCR using primers against the *GAPDH* genomic locus. Synthesis of cDNA was performed using qScript cDNA Supermix (Quanta Biosciences) from 1 µg of total RNA according to the manufacturer's protocol. cDNA was diluted 100x and 2.5 µl aliquots subjected to qPCR with SensiMix SYBR & Fluorescein Mastermix (Bioline) and appropriate primers (see Table S4) on a LightCycler 480 (Roche).

For immunofluorescence analysis, cells were seeded on coverslips and either fixed the next day for undifferentiated LUHMES cells or differentiated for defined time points and then fixed with 4% formaldehyde for 10 min at room temperature. Cells were permeabilised with 0.2% Triton X in PBS for 10 min and blocked with 10% fetal bovine serum in PBS for 30 min at RT. Primary antibody (see Table S3) incubation was performed in 1% FBS, 0.1% Tween 20 in PBS for 1 hour at RT and coverslips were washed three times with 0.1% Tween 20 in PBS. Secondary antibodies (Alexa Fluor 488, Alexa Fluor 555, Alexa Fluor 647, Thermo Scientific) were diluted 1000x in 1% FBS, 0.1% Tween 20 in PBS and applied for 1 hour at RT. Finally, cells were washed three times with 0.1% Tween 20 in PBS, stained with 5000-fold diluted DAPI in PBS for 10 min at RT, mounted using Prolong Gold or Diamond (Thermo Scientific) and dried overnight at RT. Cells were imaged on the Leica TCS Sp5 confocal microscope (Leica Microsystems) using either 40x or 63x oil immersion objectives. Image analysis was performed using Volocity (PerkinElmer). Live cell imaging of differentiating neurons was performed using the IncuCyte Zoom system (Essen BioScience) and neurite length analysis was performed using NeuroTrack package implemented into IncuCyte operating software.

Preparation of total nucleic acid for estimation of RNA versus DNA quantity

Cells were washed in PBS, lysed in lysis buffer (10mM Tris HCl [pH 7.4], 50 mM NaCl, 0.5% SDS, 100mM EDTA, 300 µg/ml proteinase K) and incubated at 50°C for 2 hours. Total nucleic acid was recovered from the completely lysed sample by ethanol precipitation in 2 volumes of 100% ethanol at room temperature (for 30 minutes), and pelleting by centrifugation. The pellet was washed once in 2 volumes of 70% ethanol, and the nucleic acid pellet was resuspended in hydrolysis buffer containing 1x DNase I buffer (*NEB*), 1mM zinc sulphate, DNase I (*NEB*) and Nuclease P1 (*Sigma*). After 4 hours, the sample was mixed thoroughly and digested for a further 8 hours. After 12 hours at 37°C, the sample was heated to 92°C for 3 minutes and cooled on ice. Two volumes of 30mM sodium acetate, 1mM zinc sulphate [pH 5.2] were added plus additional Nuclease P1 and the nucleic acids were further digested to deoxyribonucleotide and ribonucleotide 5' monophosphates for a further 24 hours at 37°C. The samples were then subjected to HPLC.

HPLC nucleotide quantifications

HPLC was performed on the 5 µm Apex ODS C18 column, with isocratic 50 mM ammonium phospate (monobasic) mobile phase. UV absorbance was recorded at 276 nm (dCMP, elution time 9.4 minutes), 282 nm (5mdCMP, elution time 17 minutes), 268 nm (dTMP, elution time 21.9 minutes), 260nm AMP and dAMP (elution times 27 minutes and 62.47 minutes) and 254 nm (GMP and dGMP, elution times 11.1 minutes and 29.7 minutes). Extinction coefficients used in nucleotide quantifications were dCMP, 8.86 x 10^{-3}; 5mdCMP 9.0 x 10^{-3}; dTMP, dGMP/GMP 12.16 x 10^{-3}; dAMP/AMP 15.04 x 10^{-3}. Relative amounts of all nucleotides were

calculated from the area under each peak (Chromeleon software) using the respective extinction coefficients.

Whole genome bisulfite seq and TAB-seq

Methylome analysis of wildtype LUHMES-derived neurons at day 9 was performed by whole genome bisulfite sequencing using an Epignome Methyl-seq kit (Epicentre, Illumina). Briefly, genomic DNA was isolated using the DNeasy Blood and Tissue kit (Qiagen) according to the manufacturer's protocol. The bisulfite reaction was performed using EZ DNA Methylation Gold kit on 100 ng of gDNA according to the manufacturer's protocol. Genomic DNA was mixed with 0.5% of unmethylated lambda DNA (Promega) for assessment of efficiency of bisulfite reaction. Random primed DNA synthesis and tagging was performed on converted gDNA. Finally, DNA was PCR amplified to generate a library for high-throughput sequencing. To map hydroxymethylcytosine in LUHMES-derived neurons, TAB-seq and ox-BS technologies were used in parallel. For TAB-seq, the TAB reaction was performed according to the published protocol (Yu et al., 2012) and genomic DNA was sonicated for 30 cycles of 30 sec ON and 30 sec OFF on low power using a Bioruptor (Diagenode). DNA was end-repaired and the ends were 3'-adenylated in order to facilitate adapter ligation. Size selection was performed using Agencourt AMPure XP (Beckman Coulter) beads and, after adapter ligation and size selection, DNA was bisulfite treated using the EpiTect Bisulfite kit (Qiagen) and PCR amplified using custom primers. All libraries were sequenced as 100 bp long pair-end reads on HiSeq 2500 Illumina platforms.

RNA-seq

Total RNA was isolated from all generated cell lines (Table S1) at day 9 of differentiation using either the RNeasy Mini kit or the AllPrep DNA/RNA Mini kit (Qiagen). Genomic DNA contamination was removed with the DNA-free kit (Ambion) and remaining DNA-free RNA was tested for purity using PCR for the *GAPDH* genomic locus. Total RNA was tested on the 2100 Bioanalyzer (Agilent Technologies) to ensure a RIN quality higher than 9 and quantified using Nanodrop. Equal amounts of total RNA (1 µg to 2.5 µg) were taken forward for library preparation and ERCC RNA Spike-in control mixes (Ambion) were added according to the manufacturer's guide. Ribosomal RNA was depleted using the Ribo-Zero Gold rRNA Removal module (Epicentre, Illumina) and isolated mRNA was tested for purity on a 2100 Bioanalyzer (Agilent Technologies). mRNA was quantified with Qubit and equal amounts used for cDNA synthesis and 3' terminal tagging using ScriptSeq v2 RNA-seq library preparation kit (Epicentre, Illumina). The library was PCR amplified to add adaptors and barcodes. RNA-seq libraries were sequenced as 100 bp pair-end reads using HiSeq 2000 or HiSeq 2500 Illumina platforms.

ATAC-seq

Chromatin accessibility in four cell lines (KO, WT, OE 4x and OE 11x, see Table S1) was quantified by ATAC-seq which was performed as in (Buenrostro et al., 2015). Neurons were scraped from the plate and nuclei were isolated using hypotonic buffer (10 mM Tris-HCl pH7.4, 10 mM NaCl, 3mM $MgCl_2$, 0.1% [v/v] Igepal CA-630) and counted. 50,000 nuclei were

resuspended in 50 µl transposition reaction mix containing 2.5 µl Nextera Tn5 Transposase and 2x TD Nextera reaction buffer. The mix was incubated for 30 min at 37 ºC. DNA was purified by either the MinElute PCR kit (Qiagen) or the Agencourt AMPure XP (Beckman Coulter) beads and PCR amplified with the NEBNext High Fidelity reaction mix (NEB) to generate a sequencing library. Libraries were sequenced as 75bp long pair-end reads on a HiSeq 2500 Illumina platform.

MeCP2 ChIP-seq

LUHMES-derived neurons at day 9 of differentiation with four levels of MeCP2: KO, WT, OE 4x and OE 11x (see Table S1) were crosslinked with 1% of Formaldehyde (Sigma) in the medium for 10 min at RT and quenched with 2.5 M Glycine (Sigma) for 2 min at RT. Cells were washed with PBS, scraped from the plate and centrifuged. Crosslinked nuclei were isolated using hypotonic buffer (10 mM Tris-HCl pH7.4, 10 mM NaCl, 3mM $MgCl_2$, 0.1% [v/v] Igepal CA-630) and counted using a haemocytometer. Chromatin from ~$4x10^6$ nuclei was sonicated for 20 cycles (30 sec ON and 30 sec OFF) using Bioruptor (Diagenode) on the high power setting. Crosslinked and sonicated chromatin was mixed with 60 ng of sonicated *Drosophila* chromatin (Active Motif) as a spike-in, and the mix was incubated overnight at 4 ºC with antibodies against MeCP2 (D4F3, Cell Signalling) plus spike-in antibody (Active Motif). After overnight incubation, magnetic Protein G coated beads (Thermo Scientific) were added and incubated for 4 hours at 4 ºC. Beads were washed, and chromatin was reverse-crosslinked overnight at 65 ºC. DNA was purified using the Agencourt AMPure XP (Beckman Coulter) beads. For ChIP-seq library preparation, IPs for each condition were pooled together to achieve 5 ng total DNA as a starting material. For example, 3-4 IPs were pooled together for the KO sample and 2 IPs were pooled for the WT sample. Libraries were prepared using the NEBNext Ultra II DNA library Prep kit (NEB) for both IPs and corresponding inputs. Libraries were sequenced as 75bp long pair-end reads on a HiSeq 2500 Illumina platform.

Quantification and statistical analysis

Statistical analysis

Calculation of standard deviation, standard error of mean and *t* tests for qPCR, Western blots, methylation and total RNA quantification using HPLC were performed using GraphPad Prism version 7.

Bisulfite sequencing

Trimmomatic version 0.32 (Bolger et al., 2014) was used to perform quality control on 94b paired-end reads to remove adapter sequence and poor quality bases at the ends of reads. We used Bismark version 0.10 (Krueger and Andrews, 2011) to further align and process the reads. Mapping was performed in bowtie2 mode to the human hg19 reference genome. Following alignment, reads were deduplicated and methylation values were extracted as bismark coverage and cytosine context files. For genome wide visualisations we calculated the methylation percentage at each cytosine position as (mC/C)x100 to generate bigWig files.

ChIP-seq data preparation

Trimmomatic version 0.32 (Bolger et al., 2014) was used to perform quality control on 75b paired-end reads to remove adapter sequence and poor quality bases at the ends of reads. We used bwa mem version 0.7.5 (Li, 2013) to map reads to the human hg19 reference genome. We filtered the alignments to remove reads that map to multiple locations in the genome and to blacklisted regions defined by the ENCODE project. We further removed duplicate reads with Picard version 1.107 MarkDuplicates (http://broadinstitute.github.io/picard/). To account for varying read depths we used deepTools version 2.5.1 (Ramírez et al., 2016) to create bigWig files normalised by RPKM (reads per kilobase per million reads).

Box 1. Accumulation algorithm for ChIP-seq.
The following example demonstrates the procedure used to obtain ChIP-seq enrichment profiles. Let us take the following sequence as the reference genome with feature of interest (in this case CG) marked red:

```
aaacagctagttaattttgaatcgcaggtaaacaatcgaataattttcta
```

Assume that ChIP-seq has generated the following reads:

```
              ttttgaatcgca
                   aatcgcaggta
             attttgaatcg
                      cgcaggtaa
                                caatcgaa
                                   tcgaataattt
```

Number of reads n_i covering a specific site in the genome is

```
0000000000000122222333443322221011122221111110000
```

Raw counts over a fixed-length region (here +/-5bp, in the actual algorithm this would be +/-5000bp) centred at the feature of interest are

```
223334433222
011122222111
```

Total accumulated counts c_i^{CG} versus position i relative to the feature of interest is therefore
```
234456655333
```
In the real analysis, the above sequence would be a 10kbp-long array of integers.

Enrichment profiles on mCG and other genomic features

To quantify MeCP2 occupancy on the genomic features of interest (mCG, GmC, CGC, etc.), we processed ChIP-seq data as follows. The input data for the analysis are the start and end position (s, e) of the ChIP reads. We reject reads longer than 1 kb regarding them as alignment artefacts. For each chromosome we create an array n_i of the number of instances a particular locus (a single base pair at position i in the chromosome) is covered by ChIP-seq by incrementing it by one for each ChIP read for which $s \leq i \leq e$. When all reads in a given chromosome have been processed, accumulated counts around features of interest (e.g.,

mCG or mCA) are further analysed as follows. For feature x at position j we select a region of interest of length L around it ($j \pm L/2$ bp) and accumulate the counts in an array (different for each feature x): $c_i^x \rightarrow c_i^x + n_{i+j}$, for all i such that $-\frac{L}{2} \leq i < \frac{L}{2}$. We proceed in a similar way for CG islands, TSS, and other regions (Figures 2F,G and S2I), in all cases positioning the midpoint j at the centre of the region of interest. Box 1 shows an example to illustrate the algorithm.

The obtained accumulated counts c_i^x for ChIP-seq are normalized (divided) by the accumulated counts for the corresponding input DNA-seq. This reduces the sequencing bias present in the data. We then calculate enrichment profiles (Figures 2C,D and S2F,G) as

$$f_i = \frac{\mathrm{Norm}_1\left(c^{ChIP,x}\right)[i]}{\mathrm{Norm}_1(c^{input,x})[i]} - 1,$$

where $\mathrm{Norm}_1(c)[i]$ normalizes the counts profiles such that their flanks have values close to one:

$$\mathrm{Norm}_1(c)[i] = \frac{c_i}{(\sum_{j=-500}^{-301} c_j + \sum_{j=301}^{500} c_j)/400}.$$

WGBS methylation data was used to determine whether a C was classified as methylated. If not stated otherwise, we assumed that at least 50% of the reads (coverage >0) must have been methylated to count a particular C as methylated.

Computer simulations of ChIP-seq

To understand how the position and the shape of the ChIP-seq enrichment profiles depend on the density of DNA-bound MeCP2, we performed computer simulations aimed at reproducing the experimental data.

5'—3' strand complementary strand	CGA GCT	CGC GCG	CGG GCC	CGT GCA	CAC GTG	CAT GTC
binding affinity relative to mCGC	50/81	1	1	1	50/57	50/81

Table S5. Relative binding affinities for different motifs. Methylated C is highlighted in red.

We assume that MeCP2 occupies methylated cytosines with probability p times the probability of binding to a particular motif (Table S5). Binding probabilities for different motifs are based on binding affinities from (Sperlazza et al., 2017) and relative binding strengths from (Lagger et al., 2017).

We then create simulated ChIP fragments (Figure 2A). Our main assumption is that if a DNA fragment (s, e) (s=start position, e=end position) contains at least one MeCP2 bound to it, it will be present in the simulated ChIP-seq. On the other hand, if the fragment does not contain any MeCP2, it may still be present in the ChIP-seq data with probability p_{bg} which accounts for "background" reads in ChIP-seq even in the absence of MeCP2. This is similar to previous models of ChIP-seq (Xu et al., 2010); even best ChIP-seq libraries can have a significant level of noise (p_{bg} close to 1) (Liang and Keles, 2012).

The input to the simulation are the *.bed files with ChIP or input (DNA) reads, and the two parameters p, p_{bg}. The simulation uses the data files to preserve the distribution of the lengths of reads but not their positions. The positions are determined by the following algorithm:

1) For each input file, GC content- and length bias is first estimated by constructing a table of counts $r[\%GC][l]$ for the actual reads and $g[\%GC][l]$ for genomic sequences of the same length l, %GC, and randomly selected locations.
2) For each read from the experimental *.bed file we calculate its length $l = e - s$ and then select a random start point s (within the same chromosome as the original read) and the end point $e = s + l$.
3) GC content and the number of mCs for the simulated read is calculated. We take a particular C to be methylated with probability proportional to the fraction of methylated reads containing this site in our methylation data (WGBS).
4) If there is at least one mC within the read, an auxiliary variable P is set to 1 with probability p times the relative binding affinity of the motif to which this C belongs. Otherwise P is set to p_{bg}. This accounts for MeCP2 present/absent in this particular DNA fragment.
5) P is multiplied by $r[\%GC][l]/g[\%GC][l]$ to account for the GC content and read length bias.
6) The simulated read is accepted with probability P and saved to a file, or rejected (with probability $1 - P$). If the read is rejected, another random position s is selected and the algorithm continues from (3).
7) Go to step 2 unless all reads from the *.bed file have been processed.

Simulated reads are then processed in the same way as the experimental ChIP data.
Since mCG is much more frequent than any other MeCP2-binding motif, in what follows we focus on enrichment on and around mCG. Figure S2C shows the enrichment profile on mCG obtained in simulations for $p_{bg} = 0$ and different p. Counterintuitively, the height of the central peak does not decrease with decreasing p, but it slowly increases. The height of the peak tends to a constant as $p \to 0$ (Figure S2D). This can be explained as follows. For large p, fragments centred at adjacent mCGs overlap, increasing the counts c_i^{mCG} in the flanking regions ($|i| \gg 1$). Normalization $\mathrm{Norm}(c)[i]$ lowers the apparent height of the peak since it divides the profile by the counts in the flanking regions. As p decreases, fragments from neighbouring mCGs overlap less; this decreases the counts in the flanks and increases the relative height of the central peak after normalization.

The overlap between fragments from neighbouring mCGs causes the shape of the peak to slightly vary with p. This can be used to approximately estimate p from the experimental data. To do this, for each ChIP-seq data set we fitted the simulated profile f_i^{sim} (parametrized by p, p_{bg}) to the experimental enrichment profile f_i^{expr} by minimizing the sum of squared differences,

$$\chi^2 = \sum_{i=-100}^{100} \left(f_i^{expr} - f_i^{sim}\right)^2,$$

with respect to p and p_{bg}. The region +/-100bp was chosen because of the sensitivity of the profile shape to p, p_{bg} in this region. Minimization was performed as follows: we simulated profiles for a range of p and two values of p_{bg}: 0 and 0.9; profiles for p_{bg} between these two values were obtained by linear interpolation (permitted due to the assumed additive model of signal and background reads). The minimum χ^2 obtained in this way for 11x OE is plotted in Figure S2E. Any $p \leq 0.1$ gives similar (low) values of χ^2, indicating that $p = 0.1$ is the upper bound on mCG occupancy that can be obtained from ChIP-seq data.

Figure S2F shows predicted profiles on features other than mCG. We have used here the values of (p, p_{bg}) obtained from the mCG profiles by minimizing χ^2 with respect to p_{bg} and assuming $p = 0.1, 0.05, 0.02, 0.002$ for 11x OE, 4x OE, WT, and KO, respectively. These values are based on the best-fit $p = 0.1$ obtained for 11x OE, the other p's being a fraction of this value approximately proportional to the relative abundance of MeCP2 in a given cell line versus 11x OE. We see in Figure S2F that strong peaks on mCA and trace peaks on GT (which MeCP2 does not bind to) are well reproduced. We stress that no additional adjustable parameters have been used here (no fitting).
We also simulated a model in which MeCP2 binds only to mCG. Figure S2G shows that while this model is equally good at reproducing peaks on mCG, it does not reproduce the peak on mCA.

ATAC-seq data preparation

Trimmomatic version 0.32 (Bolger et al., 2014) was used to perform quality control on 75b paired-end reads to remove adapter sequence and poor quality bases at the ends of reads. We used bwa mem version 0.7.5 (Li, 2013) to map reads to the human hg19 reference genome. We filtered the alignments to remove reads that map to multiple locations in the genome, to blacklisted regions defined by the ENCODE project or to mitochondrial DNA. To account for varying read depths we used deepTools version 2.5.1 (Ramírez et al., 2016) to create bigWig files normalised by RPKM (reads per kilobase per million reads).

Analysis of ATAC-seq data

ATAC-seq was analysed in a similar way to ChIP-seq, except that instead of taking the start and end positions of reads we used positions of insertion sites obtained from raw sequencing data. The positions of insertions were accumulated as in the ChIP-seq analysis to create insertion counts profiles centred at different genomic features (mCG, mCA, GT, TA). To calculate the relative insertion probability profiles showing molecular footprints of MeCP2, we calculated

$$f_i = \ln\left[\frac{\mathrm{Norm}_2\left(n_i^{cell\ line}\right)}{\mathrm{Norm}_2\left(n_i^{KO}\right)}\right],$$

where $n_i^{cell\ line}$ and n_i^{KO} are the insertion counts profiles for a given cell line and KO1, respectively, and Norm_2 normalizes the counts profiles such that their flanks have values close to one:

$$\mathrm{Norm}_2(n_i) = \frac{n_i}{\left(\sum_{j=-500}^{-301} n_j + \sum_{j=301}^{500} n_j\right)/400}.$$

ATAC-seq computer simulations

In order to understand how different sources of bias introduced by the ATAC-seq protocol affect the observed footprint of MeCP2, we developed a computer model of ATAC-seq. We use the same binding model as in the ChIP simulations (binding to mC with motif-dependent affinity). We simulate binding to a short DNA sequence (first L=50Mbp of chromosome 1). We assume that MeCP2 occupies 11bp [Nan et al., 1993] and that the protein is centred on a mC, thus the obscured genomic sequence is xxxxxmCxxxxx where x can be any nucleotide. We simulate the action of the Tn5 transposase by splitting the sequence into fragments in areas

free of MeCP2. The following algorithm (Figure 3A) is repeated T times (larger T corresponds to longer digestion times):

1) A random location i is chosen as the position of the new cut. We assume the following convention: i refers to the position of the nucleotide immediately to the left from the centre of the cut. Hence, $i+1$ denotes the position of the nucleotide to the right from the cut. Positions $i, i+1$ are the positions of insertion sites (see the above Analysis of ATAC-seq).
2) The position is accepted with probability p_i which depends on the nucleotide sequence $i-w, \ldots, i+w$ where w=10 bp. This is based on the known Tn5 sequence preference across $21 = 2w+1$ bp that it contacts (Buenrostro et al., 2013; Schep et al., 2015). The probability p_i is calculated using the position weight matrix (PWM) obtained from ATAC-seq reads for KO1.
3) If the position is rejected, go back to 1.
4) If there is enough space for Tn5, a cut is made between i and $i+1$. We assume that Tn5 occupies 21bp ($i-w, \ldots, i+w$) and for a cut to be made there must not be any protein (MeCP2) overlapping with this region. There must not be any previous cuts made in this region too.
5) If the proposed cut is rejected, go to step 1 and try again.
6) A weight $W = \exp(b \times GC)$ is assigned to each fragment where GC is the GC content (0…1) of the fragment and b is a constant. This simulates an additional GC bias that may be created by the sequencing procedure.

The simulation creates artificial fragments and their weights that we process in the same way as in the ATAC-seq analysis to obtain the simulated footprint of MeCP2. The shortest possible fragment generated by this algorithm will be thus 21bp which corresponds to two Tn5 cutting just next to each other. Since MeCP2 occupies ± 5bp, the shortest MeCP2-containing fragment is 11+21=32bp. We take only fragments >35bp since only such fragments are present in our experimental data.
The input to the program is the DNA sequence (fixed), the density p of MeCP2 on mCxx, the average density of insertion (cut) sites $t = T/L$ (cannot be larger than 1/21bp), and the GC bias b.

To test the role of the parameters of the shape and depth of the footprint of MeCP2, we simulated the model for a range of p, t, b. We also performed simulations with/without Tn5 bias.

Figure S3B shows the counts profiles for $p = 0$ (no MeCP2) with and without Tn5 bias, compared to experimental profiles. It is evident that the insertion bias is required to reproduce the experimental insertion profiles. However, the bias cancels out when calculating the relative profile (footprint) f_i. This is demonstrated in Figure S3C which shows the footprint obtained by dividing the counts profile for $p = 0.05$ by the profile for $p = 0$, for different t and $b = 6$ (Tn5 bias) or $b = 0$ (no bias). Increasing digestion time t increases the height of the side peaks surrounding a depression caused by MeCP2. However, the depth of the depression does not change noticeably.

Figure S3D, left shows that changing the GC bias b can does not change the depth of the footprint if the bias is the same in the simulated test and reference samples. However, the footprint can change if the GC bias is different in the two samples, see Figure S3D, right. Since such a mismatch in the GC bias would have a notable signature (rising/falling flanks >50bp away from mCG) which we do not see in our data, we conclude that the bias is very similar in all experimental samples.

Figure S3E shows that dividing the counts for $p > 0$ by the counts for $p = 0$ but with a different digestion time t changes the depth of the footprint slightly. Experimental variation causes t to be slightly different for different samples, and hence our results can have a small systematic error.

We also simulated the presence of nucleosomes by treating them as another protein of size 147bp that binds to the DNA, and distributing them randomly and uniformly on the DNA. We found no differences in the footprint compared to the case with no nucleosomes. However, the distribution of the number of fragments exhibited maxima at fragment lengths equal to runs of one, two, three etc. nucleosomes. This is in qualitative agreement with the ATAC-seq fragment length distribution (data not shown).

Our simulations cannot explain two subtle features of the experimental data: the presence of oscillations superimposed on the main profile, and a small central peak visible in WT/KO, CTRL/KO, and OE 4x/KO. We speculate that the first is caused by steric interactions between MeCP2 and Tn5, and the latter by interactions of proteins other than MeCP2 with mC or a small difference in the GC bias among the samples (Figure S3D, right).

Fitting simulated ATAC-seq profiles to data

The results from the previous section suggest that while some features of the footprint depends on how the samples have been prepared and sequenced (parameters t, b in our simulations), the depth of the footprint remains unaffected as long as we can guarantee that the test and control samples have been processed in a similar way.

We used the depth of the footprint to extract MeCP2 occupancy (fraction of occupied mCGs given by the parameter p). We simulated profiles for many pairs (p, t), for $p = 0 \ldots 0.1$, $t = 0 \ldots 0.08$, and a fixed $b = 6.0$ (the exact value is unimportant since GC bias cancels out when calculating f_i). We then selected those (p, t) which minimized the distance between the simulated and experimental profiles f_i^{sim} and f_i^{exp},

$$D = \sum_{i=\{-9,-10,14,15\}} \left(f_i^{exp} - f_i^{sim}\right)^2 \times \sum_{i=-100}^{100} \left(f_i^{exp} - f_i^{sim}\right)^2.$$

This formula assigns a larger weight to the region which is least sensitive to variations in the ATAC-seq protocol (edges of the central dip) and reduces the systematic error caused by the (small) central peak of unknown origin (see previous section). Figure 3B shows the best-fit profiles centred on mCG. The best fit to 11x OE footprint yields $p = 0.063, t = 0.04$. Figure 3D shows p versus MeCP2 level M for all four cell lines. The relationship is linear, with the best-fit $p = 0.0058 \times M_{cell\ line}/M_{WT}$.

Figure S3F shows profiles for OE 11x centred on other features (mCA, GT, TA). Although the fitting has been done to the mCG profile only, the footprint on mCA and the lack of it on GT, TA predicted by the simulation agree reasonably well with experimental data. This gives us confidence that the simulation captures the most important aspects of MeCP2-DNA binding and ATAC-seq. We note that the observed amplitude of the mCA footprint is slightly stronger than predicted by our simulations. A probable explanation is the reduced reliability of mCA detection in TAB-seq as compared to mCG.

ATAC-seq insertion peaks

We used the accumulated insertion counts n_i to find the positions and heights of regions of frequent insertions ("insertion peaks") as follows. For each gene, we calculated its mean insertion count $\bar{n}$ and selected regions in which $n_i > 4\bar{n}$. We then calculated the total number of insertions in these regions. Accessibility was defined as the sum of all insertions in the peaks divided by the "background" $\bar{n}$:

$$a = \frac{\sum_i n_i}{\bar{n}}.$$

RNA-seq data preparation

All paired-end sequencing reads were trimmed, and quality controlled using Trimmomatic version 0.33 (Bolger et al., 2014). The filtered reads were then mapped using STAR version 2.4.2 (Dobin et al., 2013) using hg19 human genome assembly and Ensembl 74 release for annotation. Additionally, TPM values for genomic features were quantified by quasi-mapping approach using Sailfish version 0.10.0 (Patro et al., 2014). Protein coding transcripts for Sailfish index generation were taken from Gencode release 19. Further detailed information on the parameters are in supplementary information. In order to assign reads to genomic features, featureCounts version 1.5.0 was used (Liao et al., 2014). Differential Expression analysis was performed using DESeq2 (Love et al., 2014). Comparisons between mutant vs wildtype and were performed within the same batch using the design formula ~differentiation+condition in order to account variance from the differentiation.

RNA-seq data analysis

All analysis was carried out using a subset of protein-coding genes characterized by the following features: sufficient methylation coverage (WGBS; at least 80% C detected as methylated, coverage 20 or higher), and gene bodies 1kb or longer. This resulted in 15382 genes out of the initial 17764 protein-coding genes (86%).

The variability in Log2FC for genes with similar methylation is quite high due to factors other than MeCP2. In order to get a meaningful signal representing the contribution from MeCP2 we binned the Log2FC of many genes according to their methylation. For plots of Log2FC versus methylation density (Figure 4C,E) we used bins of constant width 0.1bp/100bp. For other plots we varied the bin size to compromise between the number of plotted data points and the variability (represented by error bars in the plots) within a single bin.

To obtain the slope (Figure 4C,E) we fitted straight lines $\mathrm{Log2FC} = a\rho_{\mathrm{mCG}} + b$ to Log2FC(expression) to different ranges of $\rho_{\mathrm{mCG}} \in [\rho_1, \rho_1 + \Delta\rho]$, for different pairs $\rho_1 \in [0.5,1]$ and $\Delta\rho \in [1,4]$ (units: 1/100bp). We then selected the fit with the largest absolute slope to error ratio (largest $|a|/\sigma_a$ where σ_a is the standard error of a) and returned a of this fit as the largest computed slope of the Log2FC curve.

In all plots of Log2FC of differential gene expression (both experimental and theoretical) we also shifted the Log2FC values so that the average Log2FC in the range of mCG density $\rho_{\mathrm{mCG}} \in [1,6]$ bp/100bp was the same for all samples. This was motivated by a difficulty in determining the absolute levels of expression since we did not quantify total mRNA and the conversion from RNA-seq reads to TPMs removed any dependence on the total mRNA.

Mathematical Modelling

Generic model of gene expression. Below we list all the parameters and observables of the generic ("paradigm") model of gene expression (Shahrezaei et al., 2008) from Figure 5A which is the basis for all other models. For each gene i we define

- $k_{i,ON}, k_{i,OFF}$ – the switching rates OFF→ON and ON→OFF, respectively.
- J_i – transcription rate (in transcripts/second).
- α_i – transcription initiation rate (1/second)
- v_i - transcription elongation rate (bp/second).
- d_i – mRNA degradation rate (1/second).
- f_i – the fraction of cells in the population that at any given time have gene i in the ON state.

We assume that mRNA degrades according to 1st order kinetics (Rabani et al., 2011).The steady-state concentration c_i of mRNA from gene i is then obtained by equating the influx (through transcription in the ON state) and degradation:

$$J_i f_i = c_i d_i, \qquad \text{(Equation S1)}$$

from which it follows that

$$c_i = J_i f_i \,/ d_i. \qquad \text{(Equation S2)}$$

The number of transcripts per million bp (TPM) which we obtain from RNA-seq is thus

$$\mathrm{TPM}_i = N J_i f_i \,/ d_i, \qquad \text{(Equation S3)}$$

where N is the (generally unknown) proportionality constant that depends on the sample and is different for different cell lines, replicates and conditions but is the same for all genes in a given sample.

Condensation model. This model considers a hypothesis (ultimately proven to be false, see the main text) that MeCP2 causes chromatin condensation which reduces the fraction of cells with genes in the active (ON) state (Figure 5C). We assume that the fraction f_i of cells with gene i in the active state depends on promoter openness a_i (measured by ATAC-seq) which in turn depends on the level M of MeCP2 and gene methylation ρ_i:

$$f_i = f_i(M, \rho_i) \propto a_i = a_i(M, \rho_i). \qquad \text{(Equation S4)}$$

The model also assumes that transcription rate J_i and mRNA degradation rate d_i are not affected by MeCP2. The Log2FC of differential expression of gene i for cell lines X and Y is then

$$\mathrm{Log2FC}_{\mathrm{X/Y},i} = \log_2 \frac{\mathrm{TPM}_i(\mathrm{X})}{\mathrm{TPM}_i(\mathrm{Y})} = \log_2 \frac{N_\mathrm{X}}{N_\mathrm{Y}} + \log_2 \frac{a_i(M_\mathrm{X}, \rho_i)}{a_i(M_\mathrm{Y}, \rho_i)}.$$

The average Log2FC of genes with the same methylation density ρ is therefore

$$\mathrm{Log2FC}_{\mathrm{X/Y}}(\rho) = \log_2 \frac{N_\mathrm{X}}{N_\mathrm{Y}} + \langle \log_2 \frac{a_i(M_\mathrm{X}, \rho)}{a_i(M_\mathrm{Y}, \rho)} \rangle,$$

where $\langle \ldots \rangle$ denotes averaging over genes with the same ρ. According to the above equation, $\mathrm{Log2FC}_{\mathrm{X/Y}}$ should yield the same curve (modulo a vertical shift due to different normalization factors N_X and N_Y) as the logarithm of the ratio of accessibilities of X versus Y when plotted as a function of methylation density. Figures 5E and S5B shows that this is not the case. We can therefore reject the hypothesis that MeCP2 modulates gene expression primarily by altering the fraction of active genes.

Detachment model. In this hypothetical scenario we assume that RNA Pol II aborts transcription with some small probability λ when it collides with MeCP2 or encounters a

chemical mark left by the interaction between MeCP2 and chromatin (Figure 5F). The probability that RNA Pol II reaches the end of the gene (transcription end site, TES) is thus

$$P = (1-\lambda)^n \cong e^{-\lambda n},$$

where n is the number of "abort sites" on the gene, proportional to the number of MeCP2 molecules on the gene. ATAC-seq shows that the density of MeCP2 on a gene is proportional to its methylation density ρ and the total concentration M of MeCP2 in the nucleus, hence we can write that $n = AM\rho L = AMN_{\mathrm{mCG}}$where A is an unknown proportionality factor, L is the length of the gene, and N_{mCG} is the total number of mCGs. $\mathrm{Log2FC}$ of the differential expression X versus Y can then be written as

$$\begin{aligned}\mathrm{Log2FC}_{\mathrm{X/Y}} = \log_2 \frac{\mathrm{TPM}_i(\mathrm{X})}{\mathrm{TPM}_i(\mathrm{Y})} &= \log_2 \frac{N_{\mathrm{X}}}{N_{\mathrm{Y}}} + \log_2 \frac{P(M_{\mathrm{X}}, N_{\mathrm{mCG}})}{P(M_{\mathrm{Y}}, N_{\mathrm{mCG}})} \\ &= \mathrm{const} + \log_2 \frac{\exp(-\lambda A M_{\mathrm{X}} N_{\mathrm{mCG}})}{\exp(-\lambda A M_{\mathrm{Y}} N_{\mathrm{mCG}})} \\ &= \mathrm{const} - \gamma\ \Delta M_{\mathrm{X/Y}} N_{\mathrm{mCG}},\end{aligned}$$

where $\gamma = \lambda A M_{\mathrm{Y}} / \log_2 e$ is an unknown parameter identical for all cell lines, and $\Delta M_{\mathrm{X/Y}} = \frac{M_{\mathrm{X}}}{M_{\mathrm{Y}}} - 1$ is the relative difference between the level of MeCP2 in cell lines X and Y. For example, $\Delta M_{\mathrm{KO/WT}} = -1$ and $\Delta M_{11\mathrm{x\ OE/WT}} = 10$. $\mathrm{Log2FC}_{\mathrm{X/Y}}$ should therefore follow a straight line when plotted versus N_{mCG}, and the slope of this line should be positive for KO/WT, and negative (and 10x more steep) for 11xOE/WT. Figure 5G shows that when the model is fitted to KO data to fix the unknown constant γ, it fails to reproduce the OE 11x data. We hence conclude that there is no evidence that MeCP2 cause premature termination of transcription.

Congestion model

General considerations. We consider a hypothesis that MeCP2 slows down the elongation step of transcription by creating queues of RNA Pol II in front of MeCP2 or chemical modifications left by it in gene bodies (Figure 6A). We assume that the density of obstacles, or "slow sites", is equal to the density of molecules of MeCP2 bound to the DNA.

We show in the main text that (i) the density m of MeCP2 on the DNA is proportional to the concentration M of MeCP2 in the cell and (ii) that m is proportional to mCG density ρ. We can thus write that the transcription rate J averaged over many genes with the same mCG density ρ is

$$J = J(M\rho, \rho).$$

The second argument of J accounts for non-MeCP2 but mCG-density dependent modulation. Figure S5G show that gene expression (TPMs) versus gene body methylation for KO and 11x OE. The difference between the two cells lines is very small, and non-MeCP2 dependent component of J dominates the behaviour of $J(\rho)$. We are thus permitted to rewrite J in the factorized form

$$J \approx [1 - \epsilon(M\rho)]K(\rho),$$

where K is a fast-changing function of ρ and the correction $\epsilon(m)$ due to MeCP2 is a slowly increasing function of m, and $\epsilon(0) = 0$. This leads to the following expression for the Log2FC of cell lines X versus Y:

$$\begin{aligned}\mathrm{Log2FC}_{\mathrm{X/Y}}(\rho) = \log_2 \frac{\mathrm{TPM}_i(\mathrm{X})}{\mathrm{TPM}_i(\mathrm{Y})} &\approx \log_2 \frac{N_{\mathrm{X}}}{N_{\mathrm{Y}}} + \log_2 \frac{[1-\epsilon(M_{\mathrm{X}}\,\rho)]K(\rho)}{[1-\epsilon(M_{\mathrm{Y}}\rho)]K(\rho)} \\ &= \mathrm{const}(\mathrm{X},\mathrm{Y}) + \log_2[1-\epsilon(\rho M_{\mathrm{X}})] - \log_2[1-\epsilon(\rho M_{\mathrm{Y}}) \\ &\approx \mathrm{const}(\mathrm{X},\mathrm{Y}) - \epsilon(\rho M_{\mathrm{X}}) + \epsilon(\rho M_{\mathrm{Y}}).\end{aligned}$$

The latter approximation is valid since ϵ is supposed to be small. In particular, the Log2FC of any cell line X versus KO reads

$$\mathrm{Log2FC}_{\mathrm{X/KO}}(\rho) \approx \mathrm{const}(\mathrm{X},\mathrm{KO}) - \epsilon(\rho M_{\mathrm{X}})$$

because $\epsilon(\rho M_{\mathrm{KO}}) = \epsilon(0) = 0$. The Log2FC curves for different cell lines versus KO will therefore have the same shape when plotted in the variable ρM_{X}. It follows from this that the maximum slope of these lines will be proportional to M_{X}.

A similar argument holds when the reference is not the KO but a control line CTR with low but non-zero level of MeCP2, following a mild assumption that $\epsilon(m)$ is linear in m for small m and saturates for large m. We have

$$\mathrm{Log2FC}_{\mathrm{CTR/KO}}(\rho) \approx \mathrm{const}(\mathrm{CTR},\mathrm{KO}) - \epsilon(\rho M_{\mathrm{CTR}}).$$

We can calculate $\epsilon(m)$ from this equation:

$$\epsilon(m) \approx \mathrm{const}(\mathrm{CTR},\mathrm{KO}) - \mathrm{Log2FC}_{\mathrm{CTR/KO}}(m/M_{\mathrm{CTR}})$$

This enables us to express the Log2FC of any cell line X versus control CTR as

$$\mathrm{Log2FC}_{\mathrm{X/CTR}}(\rho) \approx \mathrm{const}(\mathrm{X},\mathrm{CTR}) - \epsilon(\rho M_{\mathrm{X}}) + \epsilon(\rho M_{\mathrm{CTR}}).$$

$$= \mathrm{const}(\mathrm{X},\mathrm{CTR}) + \mathrm{Log2FC}_{\mathrm{CTR/KO}}(\rho M_{\mathrm{X}}/M_{\mathrm{CTR}}) - \mathrm{Log2FC}_{\mathrm{CTR/KO}}(\rho)$$

The maximum slope will occur for $\rho = 0$. In the vicinity of this point,

$$\mathrm{Log2FC}_{\mathrm{X/CTR}}(\rho) \approx \mathrm{const}(\mathrm{X},\mathrm{CTR}) + \rho M_{CTR}\epsilon'(0)\left(\frac{M_{\mathrm{X}}}{M_{\mathrm{CTR}}} - 1\right),$$

where $\epsilon'(0)$ is the derivative of $\epsilon(m)$ at $m = 0$. The maximum slope is therefore proportional to $M_{\mathrm{X}}/M_{\mathrm{CTR}} - 1$. This is corroborated by experimental results presented in Figure 4D.

Specific examples of models that leads to congestion.

We consider two example mechanisms by which MeCP2 could affect elongation.

Slow sites model. In this model, MeCP2 causes a chromatin modification that slows down the RNA polymerase II. To mathematically model this process we use totally asymmetric simple exclusion process (TASEP) with open boundaries (Blythe and Evans, 2007; Derrida, 1998; Krug, 1991). A gene is represented as a one-dimensional chain of L sites. Each site is either occupied by a particle, representing RNA Pol II, or is empty. Particles enter the chain at site $i = 1$ with rate α (transcription initiation rate), move along the chain and exit at site $i = L$ with rate β. Sites can be either "fast" or "slow". Slow sites represent mCGs affected by the interaction with MeCP2, whereas fast sites are all other sites (methylated or not). The speed of the particles is v on fast sites and v_s on slow sites. Slow sites are randomly and uniformly distributed, and their density is $\rho_s = \rho p$ where ρ is the mCG density, and p is the probability that an mCG is occupied by MeCP2 (as in the ChIP-seq and ATAC-seq models). p is taken to be 0.063 for OE 11x, and proportionally smaller for other cell lines ($p = 0.0058 M_{\text{cell line}}/M_{\mathrm{WT}}$, see *Fitting simulated ATAC-seq profiles to data*).

Since RNA Pol II occupies about 60bp on the DNA and moves with the speed of about 60-70 bp/s (Fuchs et al., 2014), it is convenient to equate a single site of the model with a 60bp-long stretch of the actual DNA, and set the RNA Pol II speed to $v = 1$ sites per second on fast sites. We also assume $\beta = v = 1$ sec^{-1} and that RNA Pol II is not blocked from exiting the chain at the end, so that always $\alpha < \beta$.

We simulated this model for different chain lengths $L = 166, 500, 1666, 5000$ corresponding to gene lengths between 10kb and 300kb, and a range of initiation rates $\alpha \in [0.001,1]$, densities of slow sites $\rho_s \in [1/64, 8]$ (mCG density between 0.026 and 13.3 per 100bp), and slow-site velocities $v_s = 0.01, 0.02, 0.05, 0.1, 0.2$ (all rates are in 1/sec). For each set of (α, ρ_s, v_s) we first let the model to reach steady state ("thermalization step"). We then measured the flow J of particles (equivalent to the rate of transcription) through the chain. The flow strongly depends on ρ_s and only weakly (logarithmically) on the length L (Figure S6A,B).

Therefore, in what follows we fix $L = 5000$ sites, which is equivalent to the gene length of 300kb. The flow J obtained from these simulations is presented in Figure 6B as a function of α, for different mCG densities ρ, and for $p = 0.063$ (OE 11x). The flow is approximately linear in α until some critical α_c which depends on ρ_s, and saturates at $J = J_{max}$ when $\alpha > \alpha_c$.

To relate this model to the mRNA-seq differential expression data we must calculate Log2FC:

$$\mathrm{Log2FC_{X/Y}} = \log_2 \frac{J(\alpha, \rho_{\mathrm{s,X}})}{J(\alpha, \rho_{\mathrm{s,Y}})}$$

where $\rho_{\mathrm{s,X}} = \rho p_{\mathrm{X}}$, $\rho_{\mathrm{s,Y}} = \rho p_{\mathrm{Y}}$ in which ρ is mCG density and $p_{\mathrm{X}}, p_{\mathrm{Y}}$ are MeCP2 occupation probabilities for cell lines X,Y. We take $p_{\mathrm{X}} = \left(\frac{M_{\mathrm{X}}}{M_{\mathrm{OE11x}}}\right) p_{\mathrm{OE11x}} = 0.05 \left(\frac{M_{\mathrm{X}}}{M_{\mathrm{OE11x}}}\right)$, and similarly for p_{Y}. In the above expression we know all quantities except the initiation rate α.

To obtain the initiation rate we bin genes according to their methylation as described, and then, for each bin, fit Log2FC(11xOE/WT) from the above equation to the experimental Log2FC. This gives us $\alpha(\rho)$ as a function of methylation density ρ (Figure S6C, average $\alpha = 0.027$ s^{-1}), as a calibration, to test our model prediction on the Log2FC data of the other cell lines. This approach, rather than fitting initiation rates of individual genes, removes correlations due to gene-gene interactions and produces a relatively smooth curve $\alpha(\rho)$. We can then use the fitted $\alpha(\rho)$ to predict $\mathrm{Log2FC_{X/Y}}$ for other pairs of cells lines. The results are presented in Figure 6C,D and, as described in the main text, are in good agreement with experimental Log2FCs.

When the rates $\{\alpha\}$ are taken to be the same as for HeLa cells (1347 genes from (Fuchs et al., 2014) that were also present in our RNA-seq; average $\alpha = 0.05$ s^{-1}), the model still broadly agrees with the data, except for KO (Figure S6D) which is very sensitive to the exact values of α.

A model in which J is first evaluated for individual genes (defined by their ρ, L) and then Log2FC obtained by binning genes according to their mCG density ρ does not significantly improve the agreement (not shown).

Dynamical obstacles model. This model is very similar to the slow sites model with two exceptions: (i) polymerase always moves with the same speed v (no slow sites) as long as it is not blocked by other polymerases and MeCP2, (ii) MeCP2 binds and unbinds dynamically from the methylated sites (Figure S6E). We assume that unbinding occurs with rate k_u per MeCP2 molecule, whereas binding occurs with rate $k_u p$ per unoccupied mCG. MeCP2 does not bind if an mCG is occupied by MeCP2 or polymerase. The density of MeCP2 is thus $\approx p\rho$ in the absence of transcription and is the same as in the slow-sites model. The parameters of the model are: α, v, L, ρ, p and, in addition, the unbinding constant k_u.

The model behaves similarly to the slow-sites model. Figure S6F shows a plot of the flow as a function of α and ρ, for fixed $k_u = 0.04, p = 1$. Figures S6G,H show that (after fitting the initiation rates as described for the slow-site model) the model is also able to reproduce the experimental data, but the apparent fraction p of occupied mCGs must be close to 1. This suggests that MeCP2 may slow down RNA Pol II by altering chromatin structure rather than by direct steric interference as indicated in the main text.

Figure 1

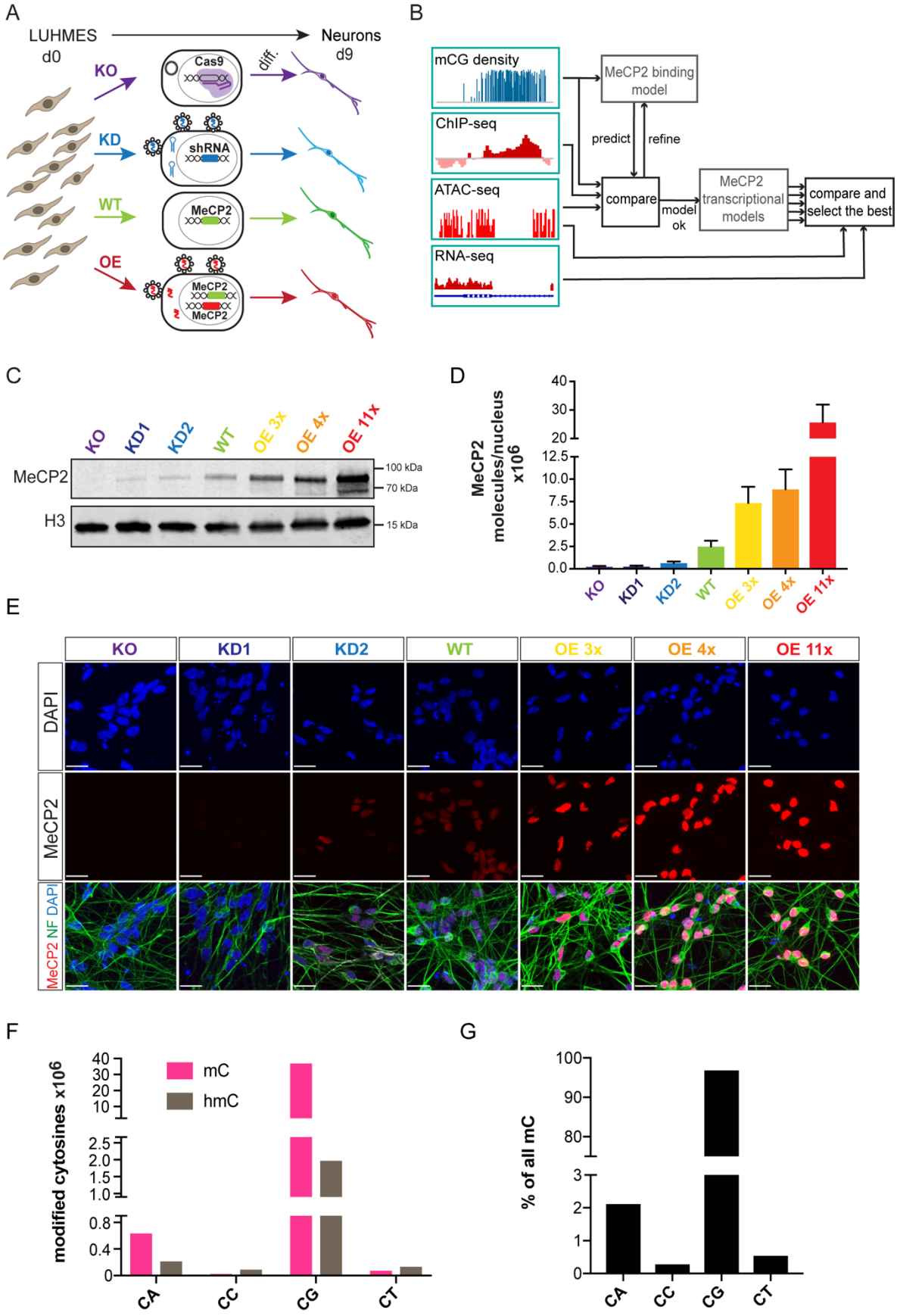

A
LUHMES
d0
Neurons
d9
KO
KD
WT
OE
Cas9
diff.
shRNA
MeCP2
MeCP2
MeCP2
B
mCG density
ChIP-seq
ATAC-seq
RNA-seq
MeCP2 binding model
predict
refine
compare
model ok
MeCP2 transcriptional models
compare and select the best
C
KO
KD1
KD2
WT
OE 3x
OE 4x
OE 11x
MeCP2
H3
100 kDa
70 kDa
15 kDa
D
MeCP2 molecules/nucleus x10^6
40
30
20
12.5
10.0
7.5
5.0
2.5
0.0
KO
KD1
KD2
WT
OE 3x
OE 4x
OE 11x
E
KO
KD1
KD2
WT
OE 3x
OE 4x
OE 11x
DAPI
MeCP2
MeCP2 NF DAPI
F
modified cytosines x10^6
mC
hmC
40
30
20
10
2.5
2.0
1.5
1.0
0.8
0.4
0
CA
CC
CG
CT
G
% of all mC
100
90
80
3
2
1
0
CA
CC
CG
CT

Figure 1. Approach used to study MeCP2 function in human neurons.

(A) Experimental design. Cell lines with different levels of MeCP2 were constructed using four different methods: CRISPR/Cas9 targeting (knock-outs KO, purple); infecting wildtype (WT, green) cells with lentiviruses encoding shRNAs against MeCP2 (knock-downs KD, blue); lentiviruses expressing MeCP2 under the control of suitable promoters (overexpressing cells OE, red).

(B) Flow diagram of the combined experimental and modelling approaches.

(C) Expression of MeCP2 in cells at day 9 of differentiation (Western blot).

(D) Mean number of MeCP2 molecules per nucleus estimated from (C). Error bars represent +/- SEM.

(E) Immunofluorescence imaging of MeCP2 and neurofilament (NF) expression in day-9 neurons. Upper panel: DAPI (blue). Middle panel: MeCP2 (red). Bottom panel: merge of upper (blue) and middle (red) panels with NF (green). Scale bar is 20 μm.

(F) Number of methylated (mC)/hydroxymethylated (hmC) dinucleotides (per haploid genome) obtained from TAB-seq in WT LUHMES-derived neurons (day 9).

(G) Percentage fraction of methylated Cs in the context of different dinucleotides calculated from TAB-seq data.

Figure 2

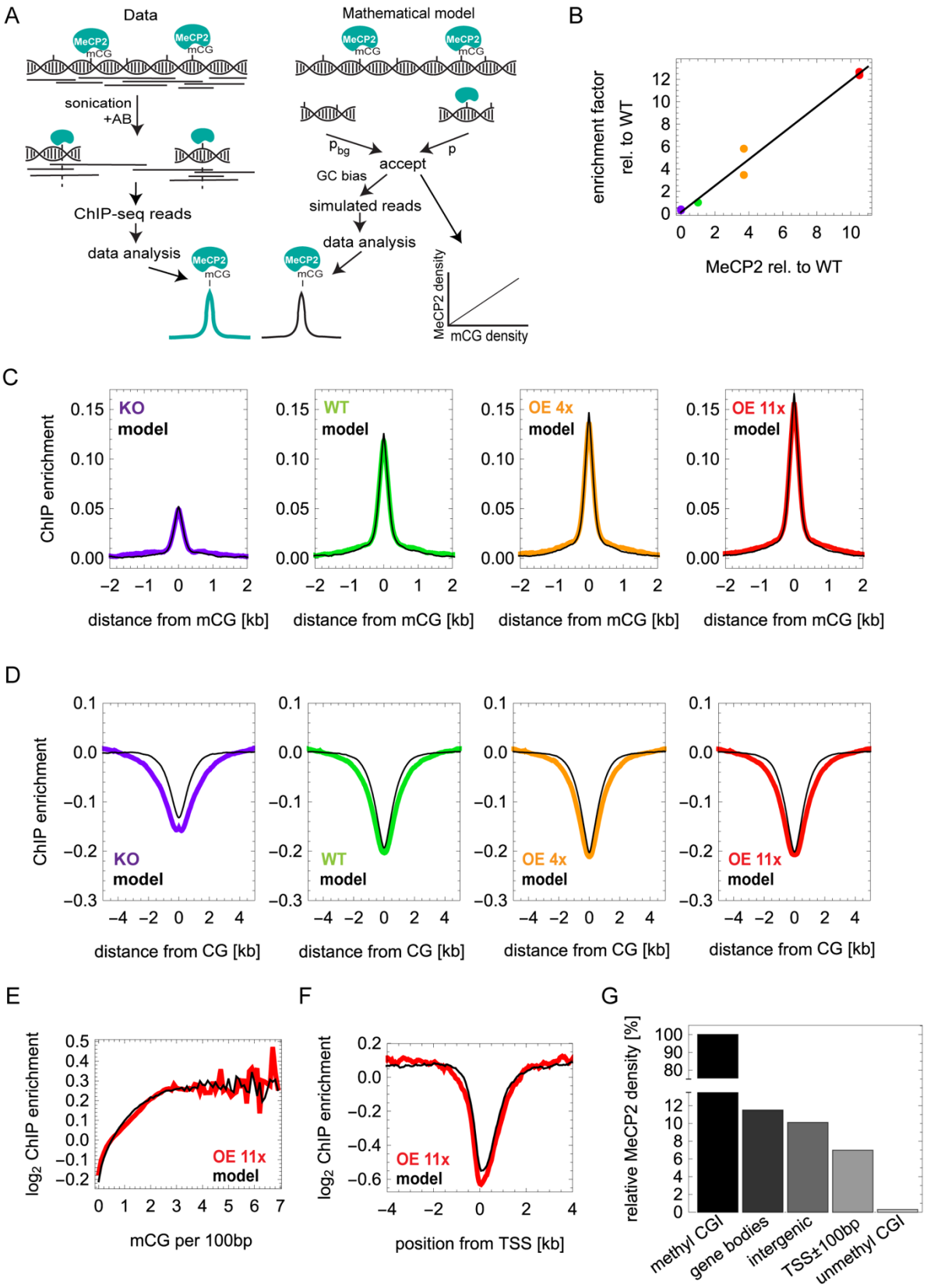

A
Data
MeCP2
mCG
sonication
+AB
ChIP-seq reads
data analysis
Mathematical model
p_{bg}
p
accept
GC bias
simulated reads
data analysis
MeCP2 density
mCG density
B
enrichment factor
rel. to WT
MeCP2 rel. to WT
C
ChIP enrichment
KO
model
WT
model
OE 4x
model
OE 11x
model
distance from mCG [kb]
D
ChIP enrichment
KO
model
WT
model
OE 4x
model
OE 11x
model
distance from CG [kb]
E
log_2 ChIP enrichment
OE 11x
model
mCG per 100bp
F
log_2 ChIP enrichment
OE 11x
model
position from TSS [kb]
G
relative MeCP2 density [%]
methyl CGI
gene bodies
intergenic
TSS±100bp
unmethyl CGI

Figure 2. Linear dependence between MeCP2 density on DNA and mCG density as revealed by quantitative ChIP-seq.

(A) MeCP2 ChIP-seq experimental procedure and its *in silico* counterpart. p_{bg} and p are probabilities of background and MeCP2-bound reads, respectively.

(B) Enrichment factor (human chromatin relative to spiked-in *Drosophila* chromatin) increases linearly with the level of MeCP2 (Fig. 1D). Two biological replicates are shown as individual data points.

(C) ChIP-seq enrichment profiles centred at mCG dinucleotides for KO (purple), WT (green), OE 4x (orange) and OE 11x (red). Black lines depict the results of the model from panel A fitted to the experimental data.

(D) As in (C) but for unmethylated CG dinucleotides. Black line = prediction of the model.

(E) Model (black line) predicts experimental ChIP-seq enrichment data in OE 11x/KO (red line) as a function of mCG density.

(F) Model (black line) predicts MeCP2 ChIP-seq enrichment profile for OE 11x/KO (red line) over all annotated TSS positions.

(G) Predicted MeCP2 density at genomic features, expressed relative to methylated CGIs (100%).

Figure 3

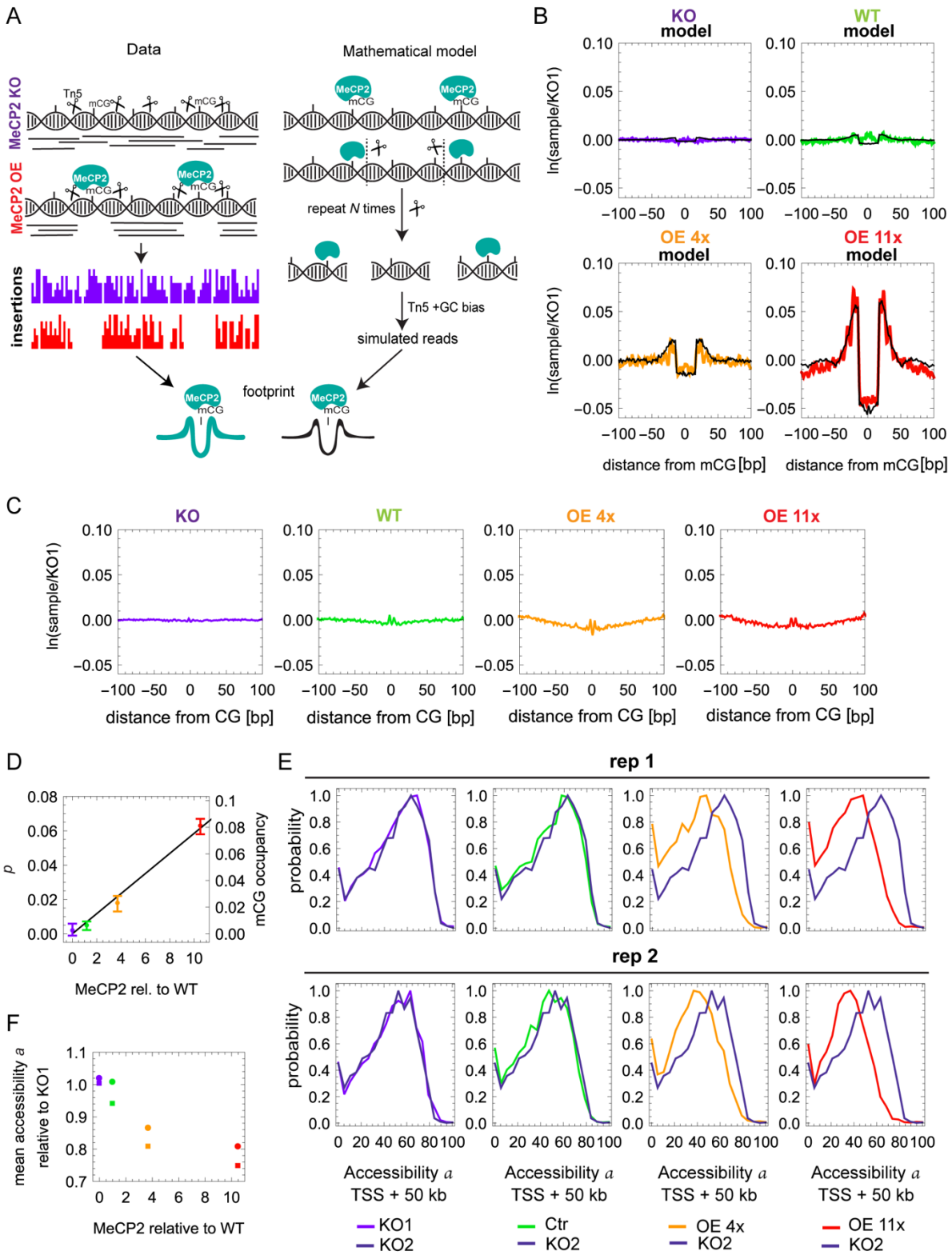

A
Data
Mathematical model
MeCP2 KO
MeCP2 OE
insertions
Tn5
mCG
MeCP2
repeat N times
Tn5 +GC bias
simulated reads
footprint
B
KO model
WT model
OE 4x model
OE 11x model
ln(sample/KO1)
distance from mCG [bp]
C
KO
WT
OE 4x
OE 11x
ln(sample/KO1)
distance from CG [bp]
D
p
mCG occupancy
MeCP2 rel. to WT
E
rep 1
rep 2
probability
Accessibility a
TSS + 50 kb
KO1
KO2
Ctr
OE 4x
OE 11x
F
mean accessibility a relative to KO1
MeCP2 relative to WT

Figure 3. ATAC-seq analysis shows MeCP2 footprint at mCG and reduced DNA accessibility in neurons over-expressing MeCP2.

(A) Schematic representation of ATAC-seq. Tn5 insertion sites (scissors) indicate DNA regions not covered by proteins (including MeCP2). A mathematical model used to infer MeCP2 occupancy on the DNA replicates all the steps of the experimental procedure.

(B) Depletion profiles (logarithm of the ratio between the number of Tn5 insertions in the sample (KO (purple), WT (green), OE 4x (orange), OE 11x (red)) and KO1) in the +/-100 bp regions surrounding mCG dinucleotides. Profiles are averages over 2-4 biological replicates. Black lines represent computer simulations of the model fitted to the data.

(C) As in (B) but for unmethylated CG dinucleotides.

(D) Predicted fraction of mCGs occupied by MeCP2 versus MeCP2 level obtained by fitting the model to the depletion profiles in (B). Error bars represent +/- SEM.

(E) Probability distribution of chromatin accessibility a in gene bodies (from the TSS to +50 kb) for four levels of MeCP2: KO (purple), control (green), OE 4x (orange), OE 11x (red). Each plot represents one biological replicate.

(F) Mean accessibility a in gene bodies (TSS to +50 kb) relative to cell line KO1, for various MeCP2 levels. Points represent two independent biological replicates (dots and squares).

Figure 4

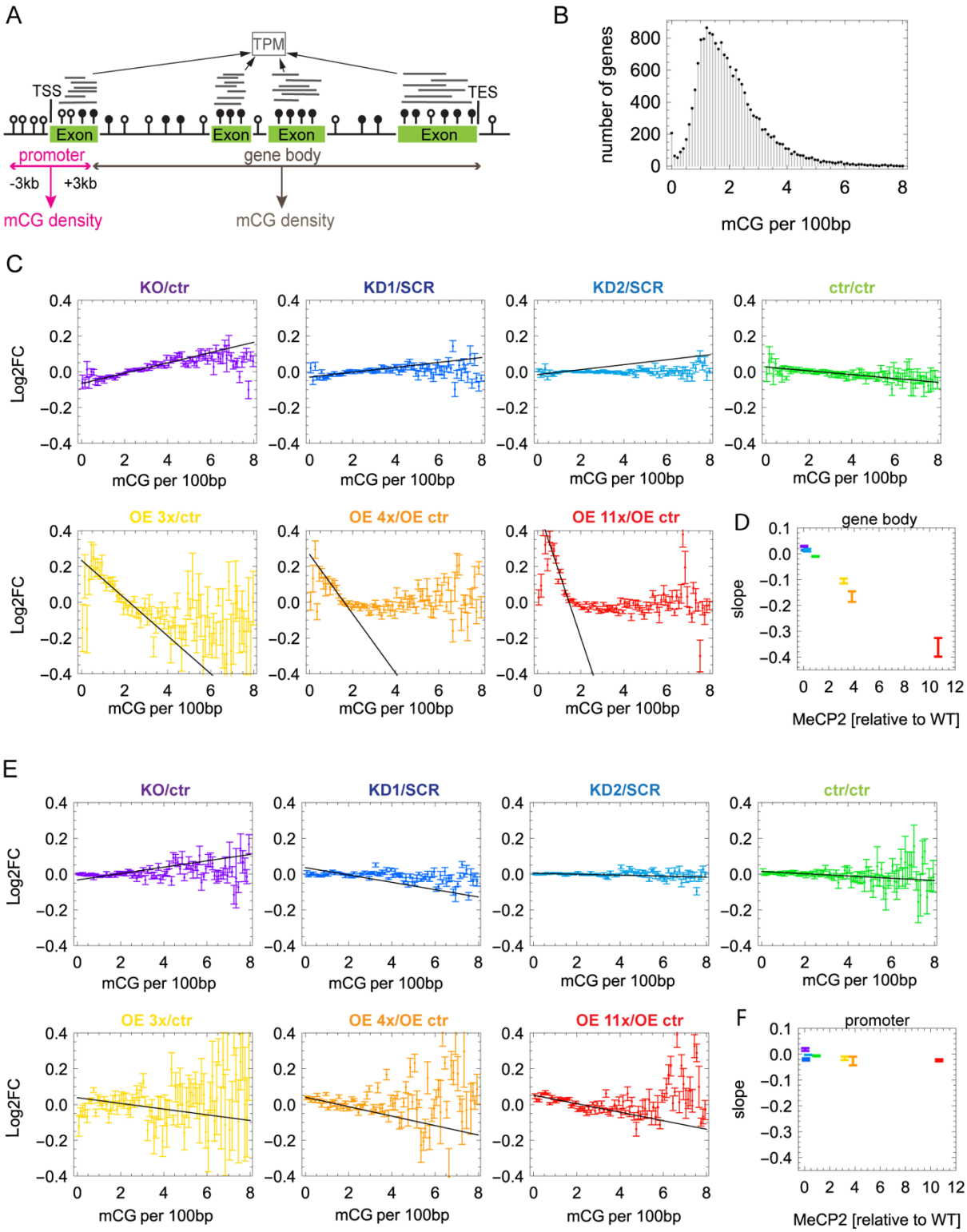

A
TPM
TSS
TES
Exon
Exon
Exon
Exon
promoter
gene body
-3kb
+3kb
mCG density
mCG density
B
number of genes
mCG per 100bp
C
KO/ctr
KD1/SCR
KD2/SCR
ctr/ctr
OE 3x/ctr
OE 4x/OE ctr
OE 11x/OE ctr
Log2FC
mCG per 100bp
D
gene body
slope
MeCP2 [relative to WT]
E
F
promoter

Figure 4. Gene expression strongly correlates with gene body mCG density and MeCP2 abundance.

(A) Schematic representation of the transcriptional unit used in this study. mCG density is the number of methylated CGs divided by gene length. TPM = transcripts per million obtained from RNA-seq. TSS = transcription start site. TES = transcription end site.

(B) Plot of all genes showing the spread of mCG density in gene bodies. mCG density for most genes is below 4/100 bp.

(C) Log2 fold change of gene expression (Log2FC) relative to appropriate controls (ctr – unmodified controls; SCR – scrambled control, OE ctr – overexpression control) for all seven levels of MeCP2, plotted against gene body mCG density. Genes were binned according to their mCG density (bin size = 0.1bp), with each point representing a mean Log2FC of all genes falling in that particular bin. Black line shows the line with the maximum slope.

(D) The maximum slope from (C) strongly correlates with the level of MeCP2 in gene bodies.

(E) As in (C) but for promoter mCG density.

(F) The slope of Log2FC (promoter mCG) shows minimal dependence on the level of MeCP2. In all panels, error bars represent +/- SEM.

Figure 5

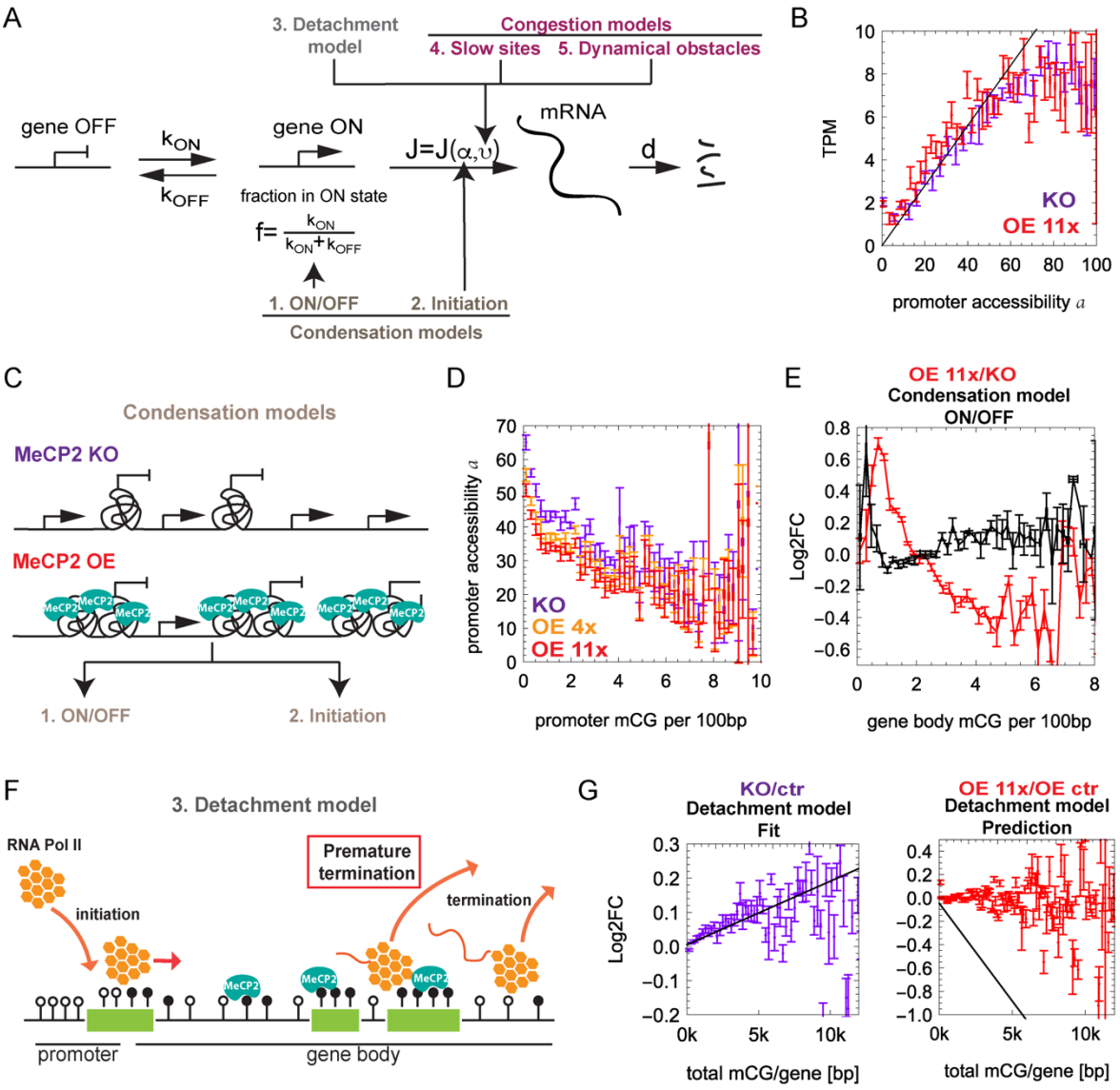

A
3. Detachment model
Congestion models
4. Slow sites
5. Dynamical obstacles
gene OFF
k_{ON}
k_{OFF}
gene ON
J=J(α,υ)
mRNA
d
fraction in ON state
$f=\frac{k_{ON}}{k_{ON}+k_{OFF}}$
1. ON/OFF
2. Initiation
Condensation models
B
TPM
KO
OE 11x
promoter accessibility a
C
Condensation models
MeCP2 KO
MeCP2 OE
1. ON/OFF
2. Initiation
D
promoter accessibility a
KO
OE 4x
OE 11x
promoter mCG per 100bp
E
OE 11x/KO
Condensation model
ON/OFF
Log2FC
gene body mCG per 100bp
F
3. Detachment model
RNA Pol II
initiation
Premature termination
termination
promoter
gene body
G
KO/ctr
Detachment model
Fit
Log2FC
total mCG/gene [bp]
OE 11x/OE ctr
Detachment model
Prediction
total mCG/gene [bp]

Figure 5. Mathematical modelling rejects hypotheses that MeCP2 regulates transcription by reducing initiation or aborting elongation of RNA polymerase II.

(A) A general model of gene transcription. Each gene can be in two states: ON (TSS accessible to Pol II) or OFF (TSS not accessible). In the ON state, mRNA is produced with rate J which depends on the transcription initiation rate α and the elongation rate v. RNA-seq does not measure J directly but rather the amount of mRNA accumulated in the cell (TPMs, transcripts per million) which also depends on degradation rate d. Five proposed models of MeCP2-dependent transcriptional regulation relate to different stages of transcription.

(B) Promoter accessibility α calculated from ATAC-seq correlates with gene expression levels (TPMs) in KO (purple) and OE 11x (red) cells as predicted by the model in (A). Black line is a linear fit to the experimental data.

(C) Schematic representation of the two Condensation models in which knots represent regions of condensed chromatin that are inaccessible to RNA Pol II, thus influencing either the ON/OFF fraction of genes (ON\OFF model 1) or transcriptional initiation around the promoter (Initiation model 2).

(D) Chromatin accessibility at promoters decreases with promoter methylation and with MeCP2 level: KO (purple), OE 4x (orange), OE 11x (red).

(E) Log2FC (gene expression) predicted by the ON/OFF Condensation model 1 (black) strongly disagrees with Log2FC(OE 11x/KO) obtained from RNA-seq (red).

(F) Schematic representation of the Detachment model 3 in which the presence of MeCP2 on the DNA causes RNA Pol II to detach and prematurely terminate transcription.

(G) Plots of Log2FC (gene expression) for: KO/ctr (left, purple) and 11x OE/OE ctr (right, red) versus the total number of mCGs in the gene. Black lines represent *in silico* predictions of the Detachment model 3. The model can be fitted to the KO/ctr data but it fails to reproduce the 11x OE/OE ctr data. Error bars represent +/- SEM.

Figure 6

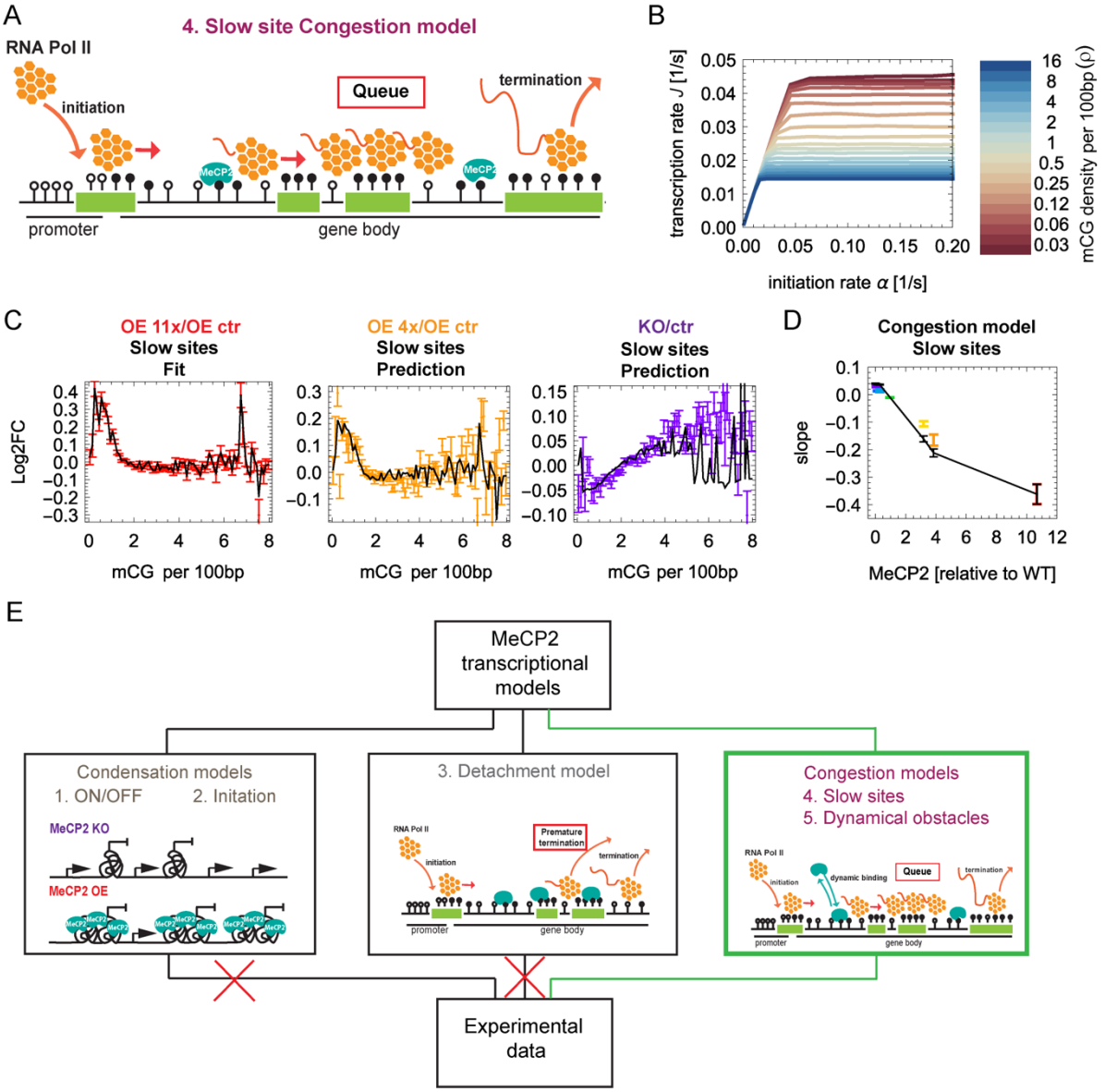

A
4. Slow site Congestion model
RNA Pol II
initiation
Queue
termination
MeCP2
promoter
gene body
B
transcription rate J [1/s]
initiation rate α [1/s]
mCG density per 100bp (ρ)
C
OE 11x/OE ctr
Slow sites
Fit
OE 4x/OE ctr
Slow sites
Prediction
KO/ctr
Slow sites
Prediction
Log2FC
mCG per 100bp
D
Congestion model
Slow sites
slope
MeCP2 [relative to WT]
E
MeCP2 transcriptional models
Condensation models
1. ON/OFF
2. Initation
MeCP2 KO
MeCP2 OE
3. Detachment model
Premature termination
Congestion models
4. Slow sites
5. Dynamical obstacles
dynamic binding
Queue
Experimental data

Figure 6. Mathematical modelling predicts that MeCP2 slows down transcriptional elongation.

(A) Schematic representation of the Slow sites Congestion model 4 in which MeCP2 slows down elongating RNA Pol II which leads to Pol II queuing.

(B) Transcription rate J predicted from model 4 (A) plotted as a function of the initiation rate α, for different gene body mCG densities. Queuing is visible as a saturation of J above some critical α that depends on mCG density. In the queuing regime, transcription rate J does not increase with increasing initiation rate α.

(C) Log2FC (gene expression) versus mCG density in gene bodies obtained in computer simulations of the Slow sites model 4 (black solid lines) fitted to the OE 11x/OE ctr RNA-seq dataset (red line) agrees well with experimental data for OE 4x/OE ctr (orange line) and KO/ctr (purple line) datasets. Error bars represent +/- SEM.

(D) The maximum slope of Log2FC (gene expression) versus mCG density in gene bodies as predicted by the Slow sites model 4 (black line) reproduces the slopes from the RNA-seq data for all seven levels of MeCP2.

(E) Summary of the results of mathematical modelling. We considered five models of MeCP2-dependent transcriptional regulation: two variants of the Condensation model (1. ON/OFF and 2. Initiation), the Detachment model (model 3) and two variants of the Congestion model (4. Slow sites and 5. Dynamical obstacles). Both Condensation and Detachment models disagree with the data, whereas the Congestion models (4. Slow sites and 5. Dynamical obstacles) are not rejected.

Figure 7

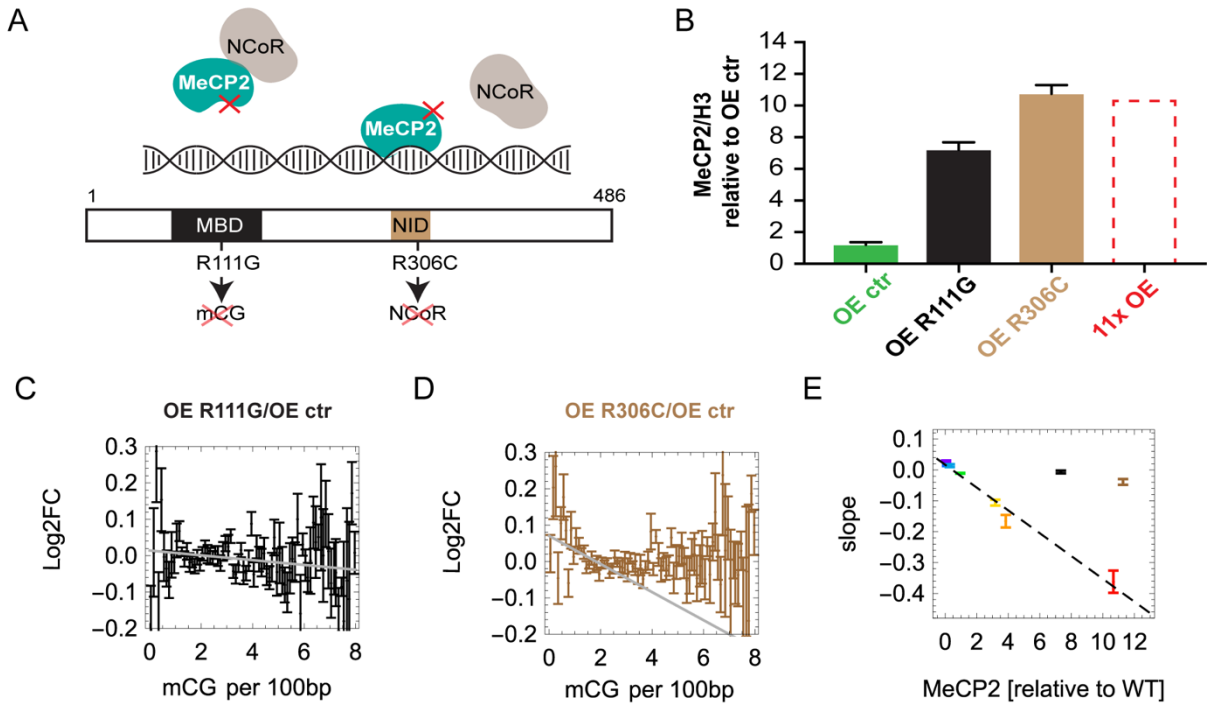


Figure 7. MeCP2 slows down transcription via a mechanism involving NCoR.

(A) Location of two binding domains in MeCP2 that are relevant for the proposed mechanism: methyl-CpG binding domain (MBD) and NCoR-interaction domain (NID). The mutation R111G causes MeCP2 to lose the ability to bind specifically to mCG. The mutation R306C prevents MeCP2 from binding the NCoR complex.

(B) Level of MeCP2 (Western blot) in two overexpressed mutant cell lines (R111G and R306C) and the overexpression control cell line (OE ctr). An outline of the OE 11x cell line is shown for comparison. Bars show the mean of three biological replicates, normalised by the H3 level.

(C) Log2FC (expression) of OE R111G/OE ctr shows almost no dependence on mCG density in gene bodies (black). Grey line shows the maximum slope. Error bars represent +/-SEM.

(D) Log2FC (expression) of OE R306C/OE ctr shows a small negative correlation with gene body mCG density (brown). Grey line shows the maximum slope.

(E) Maximum slopes for all cell lines including OE R111G (black) and OE R306C (brown) from (C and D) versus MeCP2 level (Western blot). In all plots error bars represent +/-SEM.

Figure S1_1

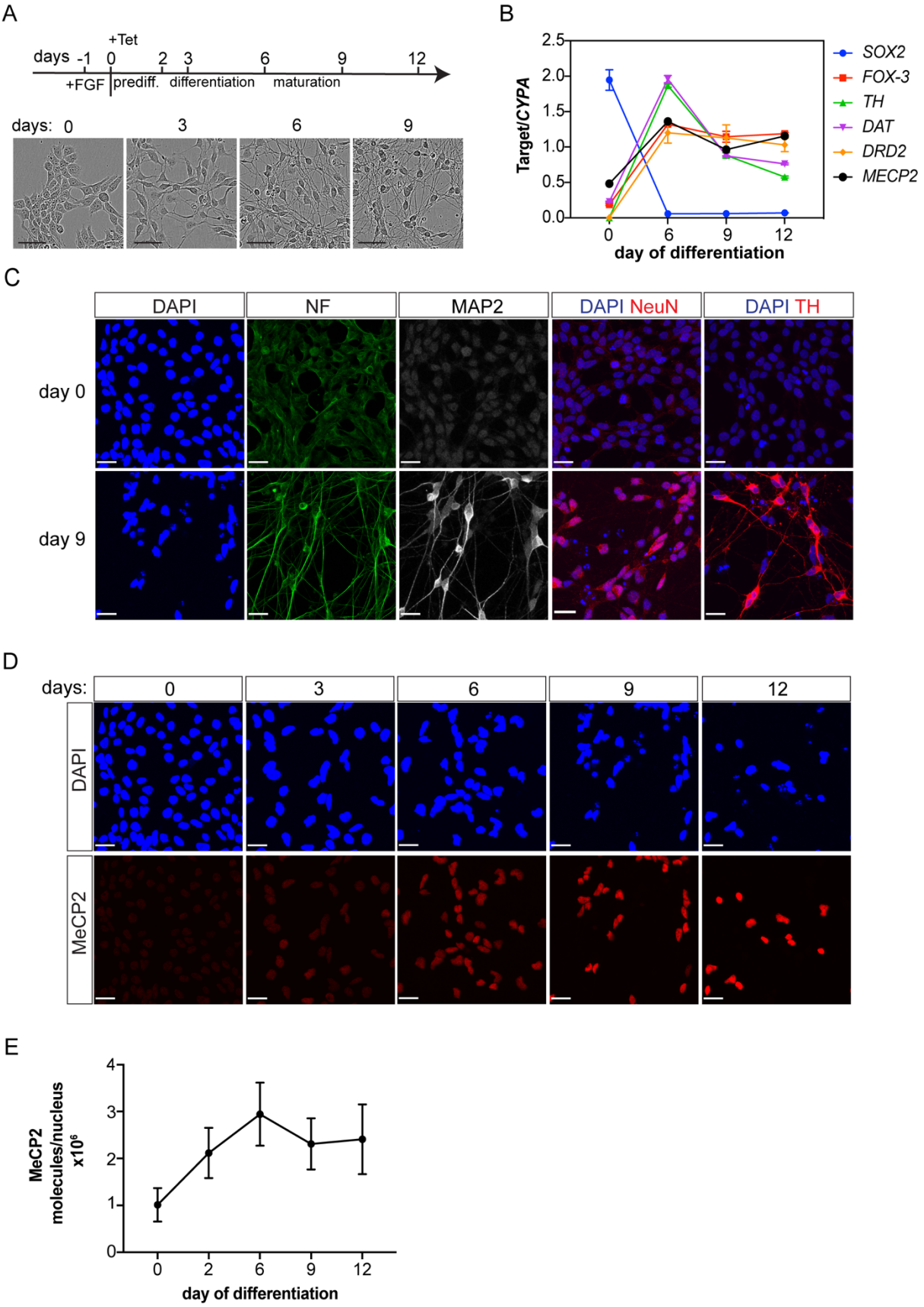

A
days -1 0 2 3 6 9 12
+Tet
+FGF
prediff.
differentiation
maturation
days: 0 3 6 9
B
Target/CYPA
day of differentiation
SOX2
FOX-3
TH
DAT
DRD2
MECP2
C
DAPI
NF
MAP2
DAPI NeuN
DAPI TH
day 0
day 9
D
days: 0 3 6 9 12
DAPI
MeCP2
E
MeCP2 molecules/nucleus x10⁶
day of differentiation

Figure S1_2

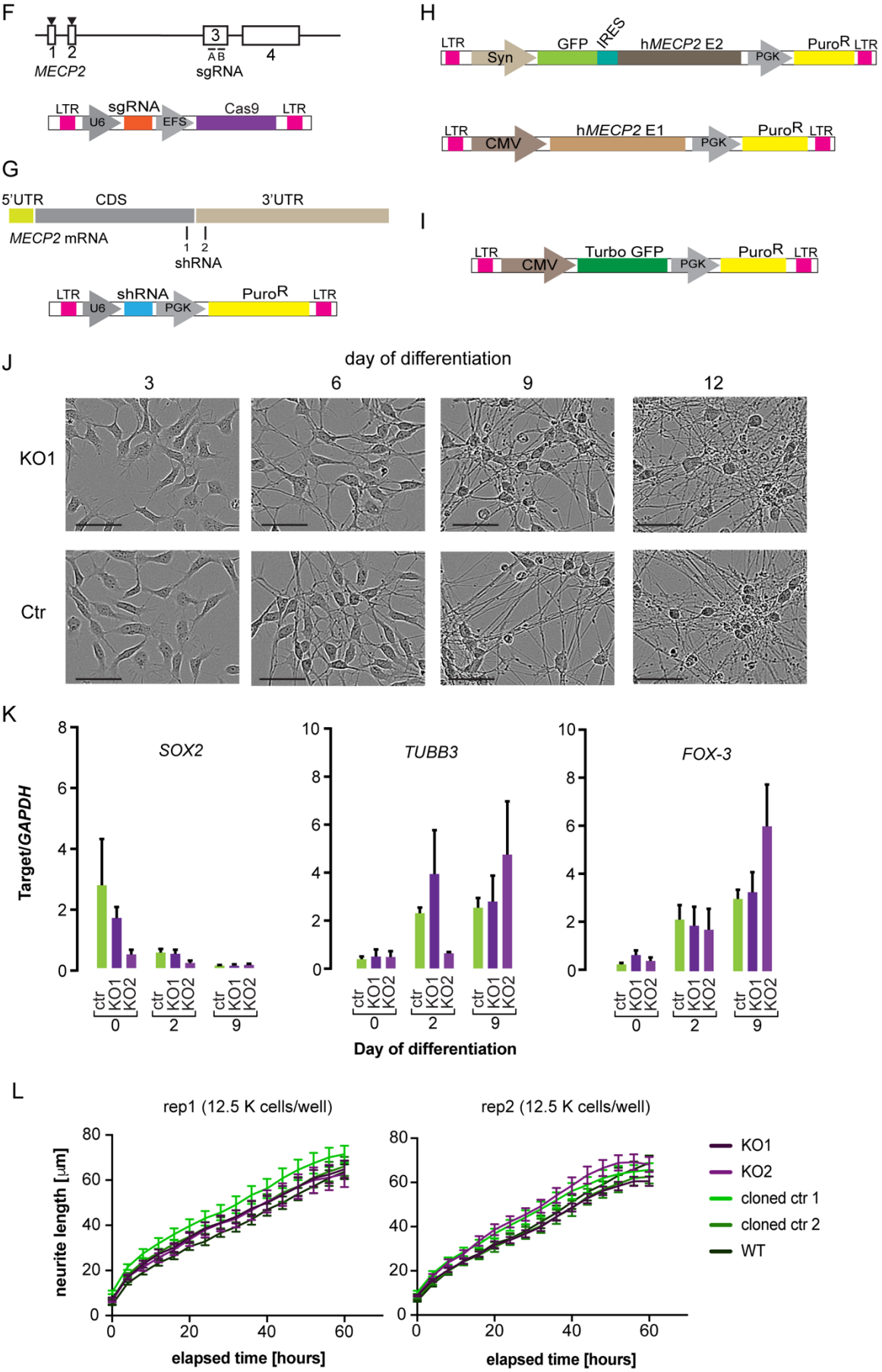

F
1 2
MECP2
3
A B
sgRNA
4
LTR
U6
sgRNA
EFS
Cas9
LTR
G
5'UTR
CDS
3'UTR
MECP2 mRNA
1 2
shRNA
LTR
U6
shRNA
PGK
PuroR
LTR
H
LTR
Syn
GFP
IRES
hMECP2 E2
PGK
PuroR
LTR
LTR
CMV
hMECP2 E1
PGK
PuroR
LTR
I
LTR
CMV
Turbo GFP
PGK
PuroR
LTR
J
day of differentiation
3
6
9
12
KO1
Ctr
K
Target/GAPDH
SOX2
TUBB3
FOX-3
ctr
KO1
KO2
0
2
9
Day of differentiation
L
rep1 (12.5 K cells/well)
rep2 (12.5 K cells/well)
neurite length [µm]
elapsed time [hours]
KO1
KO2
cloned ctr 1
cloned ctr 2
WT

Figure S1_3

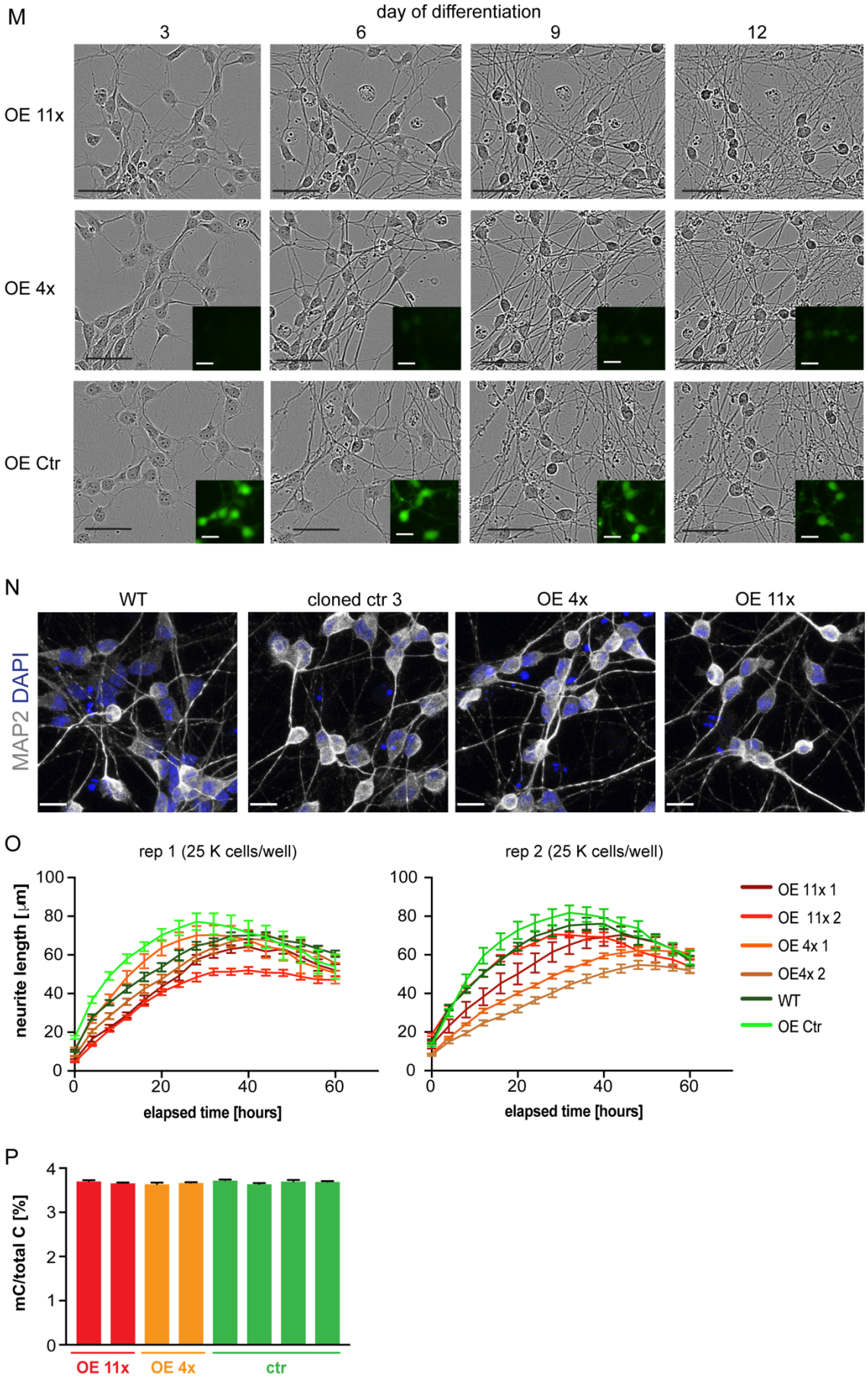

M
day of differentiation
3
6
9
12
OE 11x
OE 4x
OE Ctr
N
WT
cloned ctr 3
OE 4x
OE 11x
MAP2 DAPI
O
rep 1 (25 K cells/well)
rep 2 (25 K cells/well)
neurite length [µm]
elapsed time [hours]
OE 11x 1
OE 11x 2
OE 4x 1
OE4x 2
WT
OE Ctr
P
mC/total C [%]
OE 11x
OE 4x
ctr

Figure S1. Characterisation of LUHMES WT cells and generated cell lines. Related to Figure 1.

(A) Experimental protocol for LUHMES differentiation. Phase contrast images show cells in different stages of differentiation. Scale bar is 50 μm.
(B) Expression of neuronal differentiation markers normalized to the housekeeping gene *CYPA* during WT LUHMES differentiation. Error bars represent +/-SEM.
(C) Immunofluorescence images of undifferentiated (upper panel) and day 9 differentiated (bottom panel) WT LUHMES cells. Cells were stained with the DNA marker DAPI (blue); anti-NF (green); anti-MAP2 (grey); co-staining with DAPI plus anti-NeuN (red); co-staining with DAPI plus anti-TH (red). Scale bar is 20 μm.
(D) Immunofluorescence images showing upregulation of MeCP2 during differentiation of WT LUHMES cells (days 0, 3, 6, 9, 12). Upper panel shows DAPI staining of DNA (blue) and lower panel shows anti-MeCP2 (red). Scale bar is 20 μm.
(E) Changes in the number of MeCP2 molecules per nucleus during differentiation of WT LUHMES cells, calculated from Western blotting. Error bars represent +/-SEM.
(F) Experimental approach used to generate KO cell lines. The CRISPR system was used with one of two guide RNAs targeting common exon 3 of the *MECP2* gene which is present in all protein-coding transcripts (upper panel). Black arrowheads denote transcription start sites. Cas9 was expressed from a lentiviral plasmid transiently delivered under control of the EFS promoter.
(G) Schematic design of knock-down (KD) cell lines. Two shRNAs were designed (upper diagram) targeting the coding sequence of *MECP2* (shRNA1) and the 3'-UTR (shRNA2). The shRNAs were cloned into a lentiviral expression vector under a U6 promoter in a plasmid expressing puromycin resistance for selection.
(H) Plasmids used for stable overexpression of MeCP2. Human *MECP2* (E2 isoform) cDNA was cloned into a lentiviral plasmid under control of the human Synapsin promoter for generation of OE 3x and 4x cell lines. Additionally, OE 3x and 4x cell lines express GFP. OE 11x was obtained by cloning *MECP2* (E1 isoform) into a lentiviral plasmid under control of the CMV promoter. Both plasmids contain puromycin resistance genes to allow selection.
(I) To generate the overexpression control line (OE ctr) a lentiviral plasmid expressing Turbo GFP under the control CMV was utilised.
(J) Phase contrast images show KO1 and cloned WT control (Ctr) cells at different stages of differentiation. Scale bar 50 μm.
(K) Abundance of *SOX2*, *TUBB* and *FOX-3* mRNAs in KO1 (dark purple), KO2 (purple) and cloned control (Ctr)(green) cells as measured by qPCR, normalized to *GAPDH*. Error bars are +/-SD.
(L) Neurite length versus time from day 2-5 of differentiation (elapsed time 0-60h) for two KO cell lines (KO1 and KO2; purple), cloned controls (Ctr1 and Ctr2; green) and WT cell line (dark green). Left and right plots represent two independent differentiations (density 12.5k cells/well). Error bars are +/-SEM.
(M) Phase contrast images show OE 11x, OE 4x and OE ctr cells at different stages of differentiation. Scale bar 50 μm. Insets show GFP expression in OE control and 4x OE cells (see H and I). Scale bar for insets is 25 μm.

(N) Immunofluorescence images showing anti-MAP2 (grey) staining in WT, cloned control 3 (cloned Ctr3), OE 4x and OE 11x cells overlapped with DAPI (blue) for DNA staining. Scale bar is 20 μm.

(O) Neurite length measured for day 2-5 of differentiation (elapsed time 0-60h) for two OE 11x (OE 11x 1 and OE 11x 2; red), two OE 4x (OE 4x 1 and OE 4x 2; orange), OE control (green) and WT (dark green) cell lines. Left and right plots represent two independent differentiations (density 25k cells/well). Error bars are +/-SEM.

(P) Fraction of methylated Cs over all Cs in the genome (quantified by HPLC) of LUHMES-derived neurons at day 9 of differentiation for: OE 11x (red), OE 4x (orange) and controls (green). Error bars are +/-SEM.

Figure S2_1

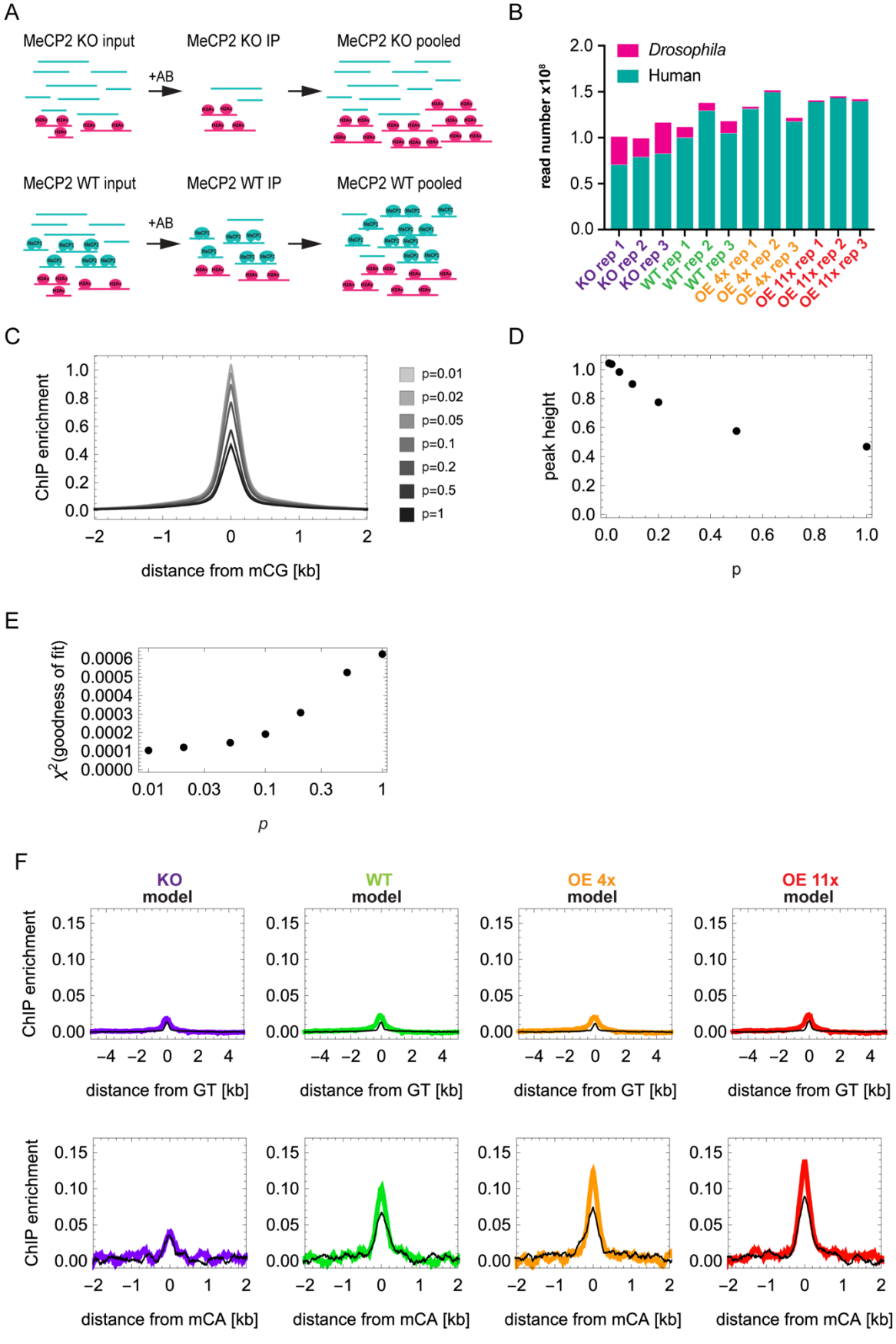

A
MeCP2 KO input
MeCP2 KO IP
MeCP2 KO pooled
+AB
MeCP2 WT input
MeCP2 WT IP
MeCP2 WT pooled
+AB
B
Drosophila
Human
read number x10^8
KO rep 1
KO rep 2
KO rep 3
WT rep 1
WT rep 2
WT rep 3
OE 4x rep 1
OE 4x rep 2
OE 4x rep 3
OE 11x rep 1
OE 11x rep 2
OE 11x rep 3
C
ChIP enrichment
distance from mCG [kb]
p=0.01
p=0.02
p=0.05
p=0.1
p=0.2
p=0.5
p=1
D
peak height
p
E
χ²(goodness of fit)
p
F
KO model
WT model
OE 4x model
OE 11x model
ChIP enrichment
distance from GT [kb]
distance from mCA [kb]

Figure S2_2

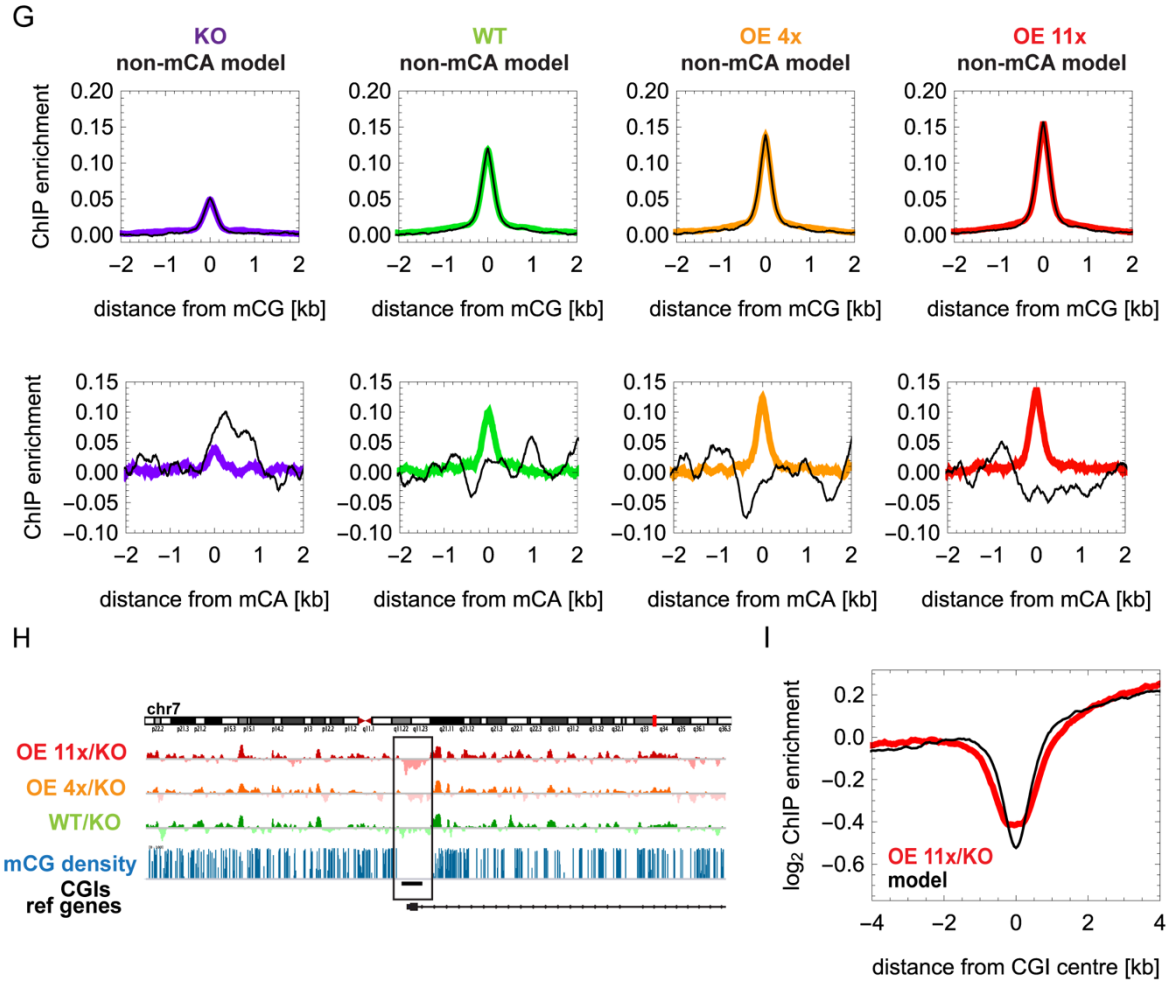

G
KO
non-mCA model
WT
non-mCA model
OE 4x
non-mCA model
OE 11x
non-mCA model
ChIP enrichment
distance from mCG [kb]
distance from mCA [kb]
H
chr7
OE 11x/KO
OE 4x/KO
WT/KO
mCG density
CGIs
ref genes
I
log2 ChIP enrichment
OE 11x/KO
model
distance from CGI centre [kb]

Figure S2. Details of the experimental and simulated ChIP-seq procedure. Related to Figure 2.

(A) Schematic representation of our quantitative MeCP2 ChIP-seq protocol for human neuronal chromatin (turquoise), using Drosophila chromatin (pink) as a spike-in for normalisation, precipitated using antibodies against H2Av.
(B) Total number of *Drosophila* reads compared to LUHMES reads obtained from ChIP-seq, for four MeCP2 levels: KO (purple), WT (green), OE 4x (orange) and OE 11x (red); each with three biological replicates.
(C) Simulated ChIP enrichment profiles for different MeCP2-DNA attachment probabilities p show a counter-intuitive reciprocal dependence between profile height and MeCP2 occupancy.
(D) The height of the peak in the enrichment profile (C) as a function of p.
(E) Goodness-of-fit $\chi 2$ as a function of p saturates below p=0.1.
(F) ChIP enrichment profiles over unmethylated GT (upper panel), and methylated CA (mCA) (bottom panel), for four levels of MeCP2. The data are well reproduced by simulations (black lines).
(G) A mathematical model (black lines) in which MeCP2 binds to mCG but is unable to bind to mCA fails to reproduce experimentally observed profiles on mCA.
(H) Representative ChIP-seq enrichment screen shot showing reads for WT (green), OE 4x (orange) and OE 11x (red) relative to KO. The mCG density track (blue) from WGBS is also shown. Black box highlights a local depletion region coincident with an unmethylated CpG island.
(I) MeCP2 ChIP-seq enrichment for OE 11x normalized to KO over all non-methylated CGIs in the genome. Red = OE 11x, black = model.

Figure S3_1

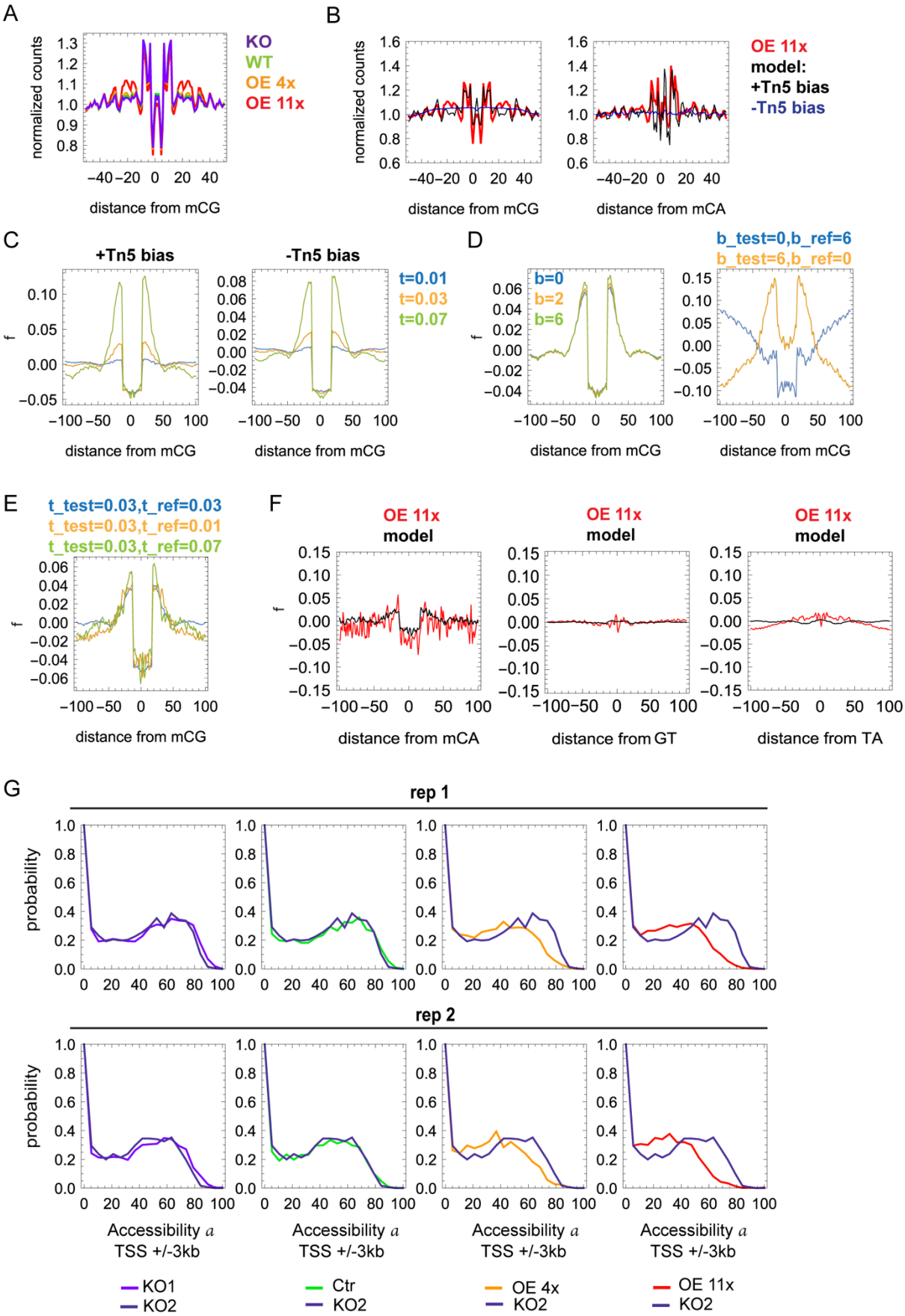

A
normalized counts
KO
WT
OE 4x
OE 11x
distance from mCG
B
normalized counts
OE 11x
model:
+Tn5 bias
-Tn5 bias
distance from mCG
distance from mCA
C
+Tn5 bias
-Tn5 bias
t=0.01
t=0.03
t=0.07
f
distance from mCG
D
b=0
b=2
b=6
b_test=0,b_ref=6
b_test=6,b_ref=0
E
t_test=0.03,t_ref=0.03
t_test=0.03,t_ref=0.01
t_test=0.03,t_ref=0.07
F
OE 11x
model
distance from mCA
distance from GT
distance from TA
G
rep 1
rep 2
probability
Accessibility a
TSS +/-3kb
KO1
KO2
Ctr
OE 4x
OE 11x

Figure S3. Details of the experimental and simulated ATAC-seq procedure. Related to Figure 3.

(A) Normalized, accumulated experimental ATAC-seq counts versus the distance from mCG for four levels of MeCP2: KO (purple), WT (green), OE 4x (orange); OE 11x (red). The visible structure is primarily due to Tn5 insertion bias. The presence of MeCP2 shows as a small deviation of the OE 11x profile from the KO profile.

(B) Comparison between experimental and simulated (p=0, t=0.04, g=6) ATAC-seq counts shows that Tn5 bias is important in reproducing the normalized counts in silico. Red = OE 11x, blue = simulated with no Tn5 bias, black = simulated with Tn5 bias.

(C) Simulated footprint f for p=0.05, g=6 and t=0.01, 0.03, 0.07 (blue, yellow, and green respectively) shows that excluding the Tn5 insertion bias does not significantly affect the depth of the footprint. Left = with Tn5 bias, right = no bias.

(D) The role of CG bias b on the simulated footprint f. CG bias cancels out if identical in both test and reference samples. Left: the same b=0,2,6 (blue, yellow, and green respectively) for the test and reference samples. The footprint is not affected by the bias. Right: different b's for the test and reference samples: b_{test}=0, b_{ref}=6 (blue) and b_{test}=6, b_{ref}=0 (yellow). The shape and depth of the footprint is significantly affected. In all cases p=0.04.

(E) The shape of the footprint depends on the difference in digestion time t between the test and the reference sample, but the depth of the footprint does not. Blue = (t_{test}=0.03, t_{ref}=0.03), yellow = (t_{test}=0.03, t_{ref}=0.01), green = (t_{test}=0.03, t_{ref}=0.07). In all cases, p=0.05, b=6.0.

(F) MeCP2 footprint on features other than mCG: mCA, GT, TA for OE 11x (red) normalized by KO, compared to predicted footprint profiles from simulations (black). As in ChIP-seq, the footprint is present on mCA and absent on GT and TA.

(G) Probability distribution of accessibility a calculated in regions near gene promoters (TSS +/-3kb) for four levels of MeCP2: KO (purple), control (green), OE 4x (orange), OE 11x (red) (one replicate for each plot). Similarly to Fig. 3E, accessibility is reduced (distributions shifted towards a=0) for OEs.

Figure S4

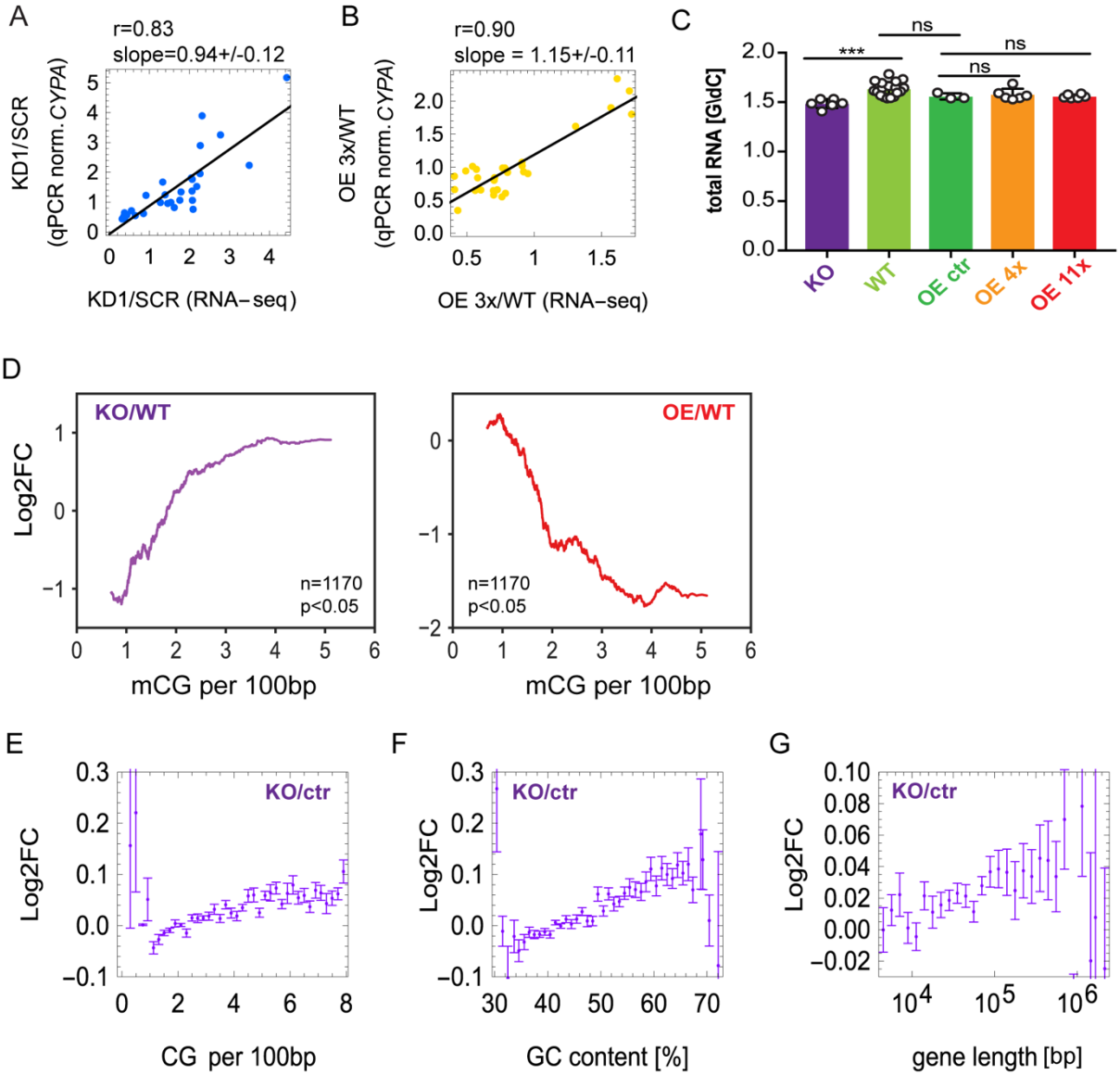

A
r=0.83
slope=0.94+/-0.12
KD1/SCR (qPCR norm. CYPA)
KD1/SCR (RNA−seq)
B
r=0.90
slope = 1.15+/-0.11
OE 3x/WT (qPCR norm. CYPA)
OE 3x/WT (RNA−seq)
C
total RNA [GldC]
***
ns
ns
ns
KO
WT
OE ctr
OE 4x
OE 11x
D
KO/WT
OE/WT
Log2FC
n=1170
p<0.05
mCG per 100bp
E
KO/ctr
Log2FC
CG per 100bp
F
KO/ctr
Log2FC
GC content [%]
G
KO/ctr
Log2FC
gene length [bp]

Figure S4. Gene expression correlates with many genomic features. Related to Figure 4.

(A) Validation of RNA-seq by qPCR normalized to *CYPA* of 29 selected genes (based on significance of Log2FC in RNA-seq) in the KD1 versus scrambled control (SCR). A good linear correlation between the RNA-seq and qPCR results is obtained.
(B) As in (A) but for OE 3x.
(C) Total RNA from day 9 neurons normalized to genomic DNA for four levels of MeCP2 expression: KO (purple), WT (green), OE control (dark green), OE 4x (orange) and OE 11x (red). The significant difference (~8%, p=0.0001) is between KO and WT. Points are individual biological replicates. Error bars represent mean +/- SD.
(D) Strong reciprocal correlation of Log2FC relative to appropriate controls for KOs (purple, left panel) and OEs (red, right panel), plotted against gene body mCG density for significantly changing genes $p<0.05$; n=1170.

(E-G) Log2FC of gene expression for KO/ctr (purple) correlates with:

(E) gene body CG density regardless of its methylation,
(F) gene body GC content,
(G) gene length.

Figure S5

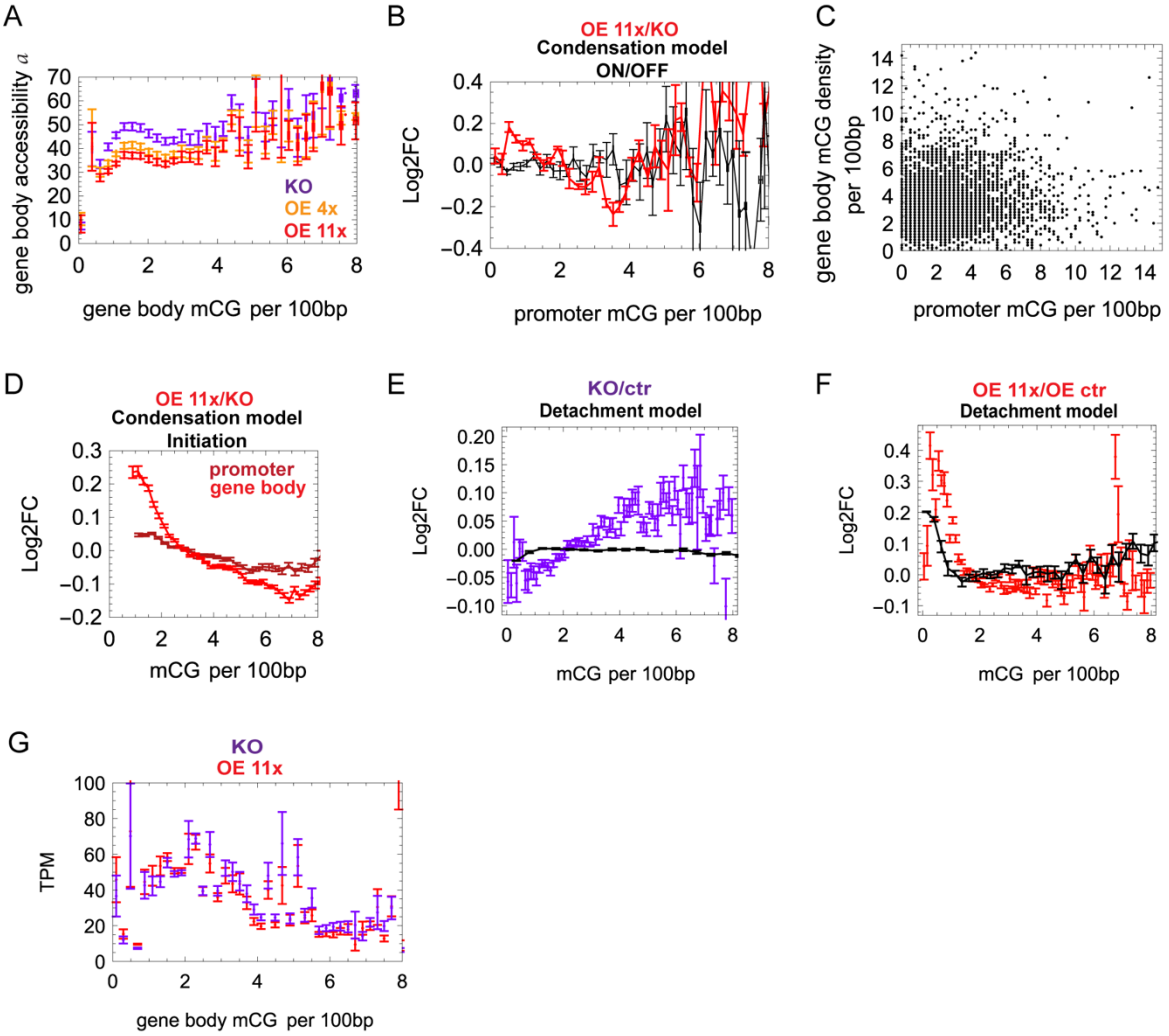

A
gene body accessibility a
KO
OE 4x
OE 11x
gene body mCG per 100bp
B
OE 11x/KO
Condensation model
ON/OFF
Log2FC
promoter mCG per 100bp
C
gene body mCG density per 100bp
promoter mCG per 100bp
D
OE 11x/KO
Condensation model
Initiation
promoter
gene body
Log2FC
mCG per 100bp
E
KO/ctr
Detachment model
Log2FC
mCG per 100bp
F
OE 11x/OE ctr
Detachment model
Log2FC
mCG per 100bp
G
KO
OE 11x
TPM
gene body mCG per 100bp

Figure S5. Condensation and Detachment models fail to reproduce experimental data. Related to Figure 5.

(A) Gene body accessibility a (ATAC-seq) varies with gene body mCG density for three MeCP2 levels: KO (purple), OE 4x (orange) and OE 11x (red).
(B) Log2FC (gene expression) predicted by the ON/OFF Condensation model (model 1) (black line) in which MeCP2 affects the fraction of genes in the ON/OFF state by binding to promoters deviates from Log2FC for OE 11x/KO obtained from RNA-seq (red) in a range of mCG density between 0.8 and 4 mCG per 100bp when plotted against promoter mCG density.
(C) Promoter mCG density (TSS+/-1 kb) does not correlate (Pearson's r=0.084) with gene body mCG density (40 kb downstream of the TSS, +/- 1 kb).
(D) Prediction of Initation model (model 2) does not agree with the experimental data. Log2FC of gene expression for 11x OE/KO correlates more strongly with gene body mCG density (calculated from +/-1kb region 40 kb downstream from TSS, bright red line) than with promoter mCG density (calculated from +/-1kb region around TSS, dark red).
(E) Log2FC (gene expression) predicted by the Detachment model (model 3) (black line) does not agree with Log2FC for KO/ctr from RNA-seq (purple) when plotted against gene body mCG density.
(F) Same as in (E) for Log2FC (gene expression) OE 11x/OE ctr (red).
(G) Plots of gene expression (transcripts per million, TPM) versus gene body mCG density for KO and OE 11x show that small differences between the two cell lines are overshadowed by a much stronger dependence on gene body mCG density independent of MeCP2.

Figure S6_1

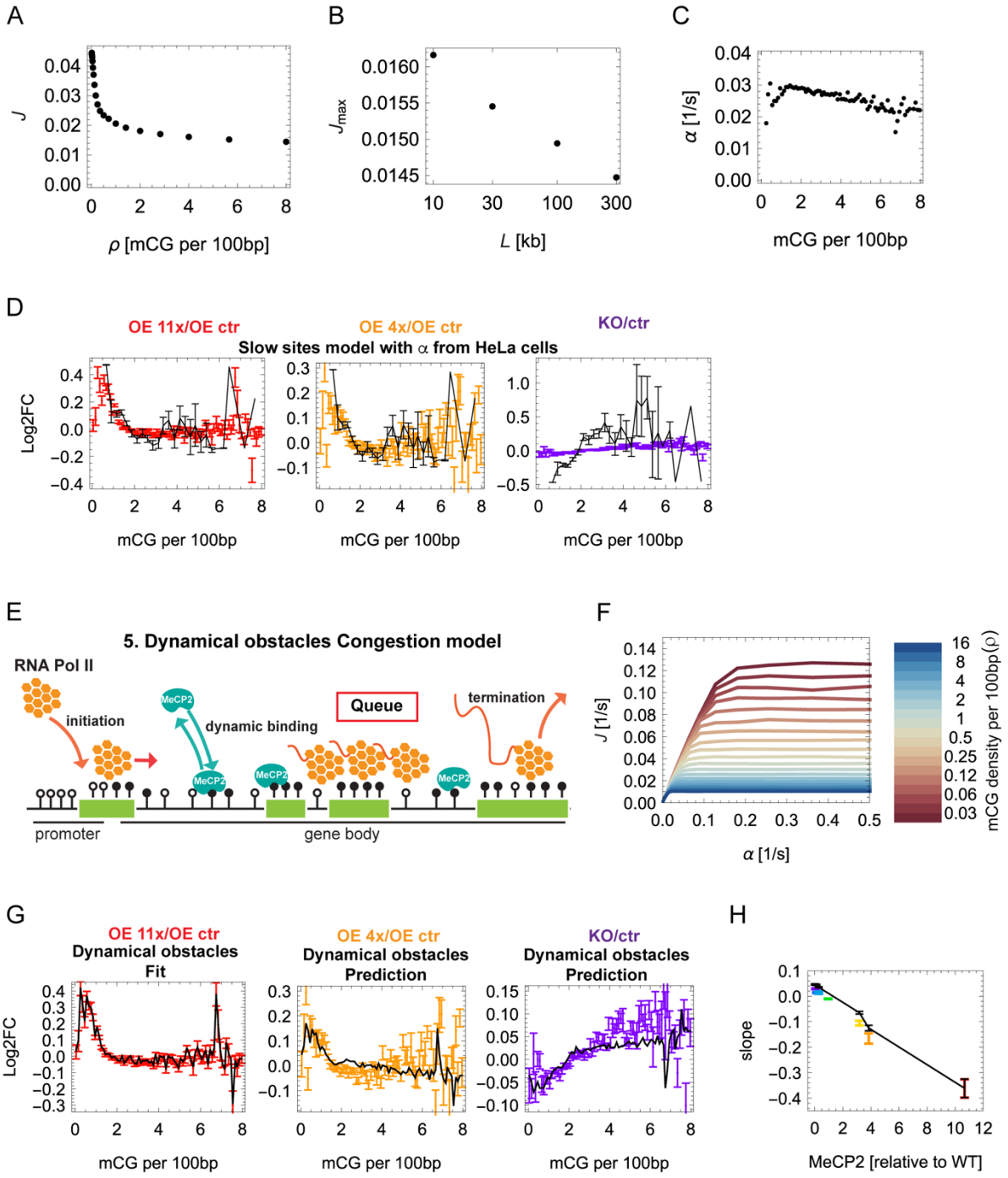

A
J
ρ [mCG per 100bp]
B
Jmax
L [kb]
C
α [1/s]
mCG per 100bp
D
OE 11x/OE ctr
OE 4x/OE ctr
KO/ctr
Slow sites model with α from HeLa cells
Log2FC
mCG per 100bp
E
5. Dynamical obstacles Congestion model
RNA Pol II
initiation
MeCP2
dynamic binding
Queue
termination
promoter
gene body
F
J [1/s]
α [1/s]
mCG density per 100bp(ρ)
G
OE 11x/OE ctr
Dynamical obstacles
Fit
OE 4x/OE ctr
Dynamical obstacles
Prediction
KO/ctr
Dynamical obstacles
Prediction
Log2FC
mCG per 100bp
H
slope
MeCP2 [relative to WT]

Figure S6. Details of the Congestion models. Related to Figure 6.

(A-D) Slow sites Congestion model (model 4):

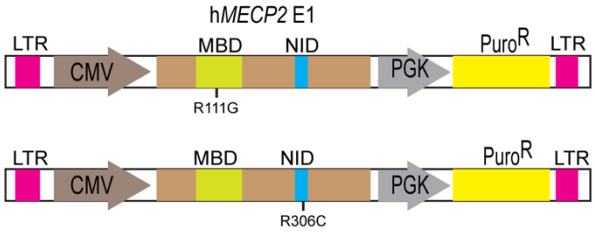
(A) Transcription rate J quickly decreases with gene body mCG density ρ. Parameters: L=5000 sites (300kb), α=0.064, p=0.063, v_S=0.05.
(B) Transcription rate J depends weakly on gene length L. Parameters: α=1, p=0.063, v_S=0.05.
(C) Initiation rate α as a function of mCG density ρ obtained by fitting the model to Log2FC(OE 11x/KO) decreases slowly with ρ. Parameters: L=5000 sites, p=0.063, v_S=0.05.
(D) Log2FC predicted by the model 4 (black line) in which initiation rates $\{\alpha_i\}$ have been assumed to be the same as for HeLa cells (Fuchs et al., 2014) versus experimental Log2FC (gene expression) for OE 11x (red), OE 4x (orange) and KO (purple). Parameters as before.

(E-H) Dynamical obstacles Congestion model (model 5):

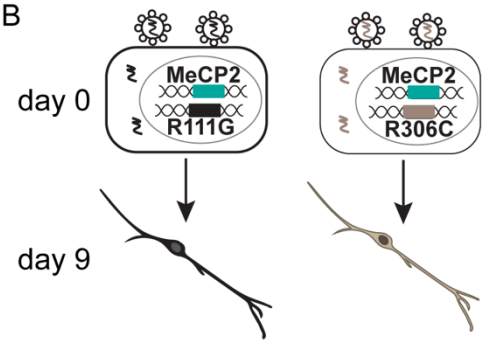
(E) Schematic representation of Dynamical obstacles Congestion model (model 5) in which dynamic biniding/unbinding of MeCP2 changes the elongation rate of RNA Pol II which in turn leads to its queuing.
(F) Transcription rate J versus initiation rate α saturates above some αC which depends on mCG density ρ. This behaviour of model 5 is similar to that of the Slow sites model (model 4) (Figure 6B).
(G) Log2FC obtained from the Dynamical obstacles Congestion model (model 5) (black line) agree well with experimental data (red, orange and purple). Left panel: model fitted to OE 11x to obtain $\alpha(\rho)$. Middle and right panel: model predictions versus experimental Log2FC for OE 4x (orange) and KO (purple). Parameters: L=5000 sites, p=1, k_u=0.04.
(H) The maximum slope of Log2FC (gene expression) versus mCG density in gene bodies as predicted by the Dynamical obstacles model (model 5) (black line) reproduces the slopes from RNA-seq data for all seven levels of MeCP2.

Figure S7

A

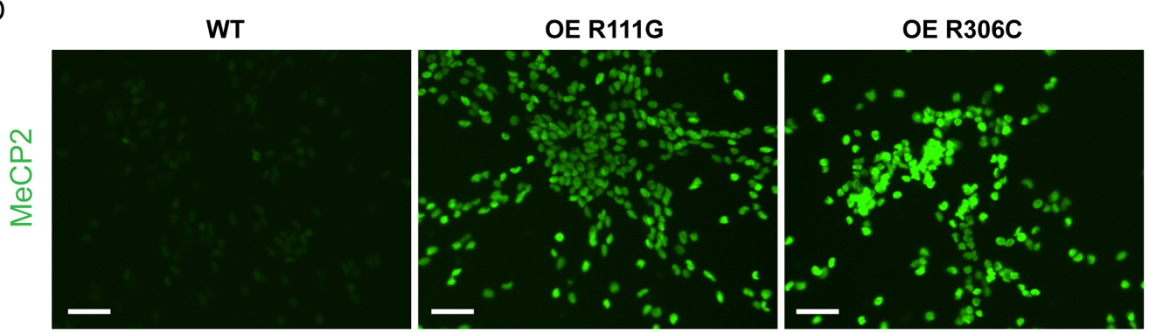


B



C

D



E

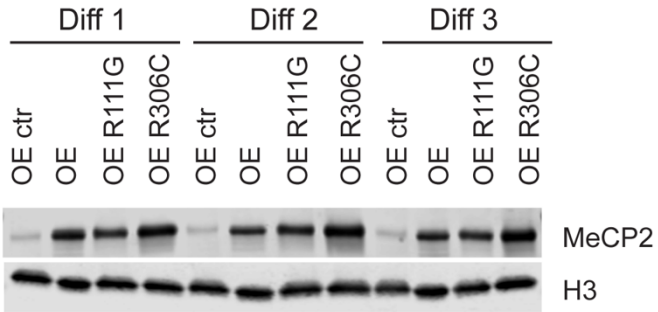


Figure S7. Generation and characterisation of over-expressing MeCP2 mutants R111G and R306C. Related to Figure 7.

(A) Plasmids used for stable overexpression of mutant MeCP2 R111G (upper panel) and R306C (bottom panel). *MECP2* (E1 isoform) was cloned into a lentiviral plasmid under control of the CMV promoter and mutated by site-directed mutagenesis. Both plasmids contain puromycin resistance genes to allow selection.

(B) Cell lines overexpressing MeCP2 with different mutations R111G (black) and R306C (brown) were generated by lentiviral transduction into undifferentiated LUHMES. After clonal selection over-expressing MeCP2 mutants were differentiated for 9 days.

(C) DNA sequencing confirms the presence of substitutions R111G and R306C following the integration of the respective constructs into LUHMES-derived neurons for two independent cell clones for each genotype.

(D) Immunofluorescence images with an antibody against MeCP2 show uniform overexpression of MeCP2 mutants R111G and R306C in the LUHMES-derived neurons at 9 days of differentiation. Scale bar is 50 μm.

(E) Overexpression of R111G and R306C compared with control cells (OE ctr) confirmed by Western blots in three independent differentiations.