



THE UNIVERSITY *of* EDINBURGH

Edinburgh Research Explorer

Quantitative modelling predicts the impact of DNA methylation on RNA polymerase II traffic

Citation for published version:

Cholewa-Waclaw, J, Shah, R, Webb, S, Chhatbar, K, Ramsahoye, B, Pusch, O, Yu, M, Greulich, P, Waclaw, B & Bird, AP 2019, 'Quantitative modelling predicts the impact of DNA methylation on RNA polymerase II traffic', *Proceedings of the National Academy of Sciences (PNAS)*, vol. 116, no. 30, pp. 14995-15000. <https://doi.org/10.1073/pnas.1903549116>

Digital Object Identifier (DOI):

[10.1073/pnas.1903549116](https://doi.org/10.1073/pnas.1903549116)

Link:

[Link to publication record in Edinburgh Research Explorer](#)

Document Version:

Peer reviewed version

Published In:

Proceedings of the National Academy of Sciences (PNAS)

General rights

Copyright for the publications made accessible via the Edinburgh Research Explorer is retained by the author(s) and / or other copyright owners and it is a condition of accessing these publications that users recognise and abide by the legal requirements associated with these rights.

Take down policy

The University of Edinburgh has made every reasonable effort to ensure that Edinburgh Research Explorer content complies with UK legislation. If you believe that the public display of this file breaches copyright please contact openaccess@ed.ac.uk providing details, and we will remove access to the work immediately and investigate your claim.



Quantitative modelling predicts the impact of DNA methylation on RNA polymerase II traffic

Justyna Cholewa-Waclaw¹, Ruth Shah¹, Shaun Webb¹, Kashyap Chhatbar¹, Bernard Ramsahoye¹, Oliver Pusch², Miao Yu³, Philip Greulich⁴, Bartlomiej Waclaw¹, Adrian P Bird¹

¹University of Edinburgh, ²Medical University of Vienna, ³Ludwig Institute for Cancer Research, La Jolla, ⁴University of Southampton

Submitted to Proceedings of the National Academy of Sciences of the United States of America

Patterns of gene expression are primarily determined by proteins that locally enhance or repress transcription. While many transcription factors target a restricted number of genes, others appear to modulate transcription levels globally. An example is MeCP2, an abundant methylated-DNA binding protein that is mutated in the neurological disorder Rett Syndrome. Despite much research, the molecular mechanism by which MeCP2 regulates gene expression is not fully resolved. Here we integrate quantitative, multi-dimensional experimental analysis and mathematical modelling to indicate that MeCP2 is a novel type of global transcriptional regulator whose binding to DNA creates "slow sites" in gene bodies. We hypothesise that waves of slowed-down RNA polymerase II formed behind these sites travel backward and indirectly affect initiation, reminiscent of defect-induced shock waves in non-equilibrium physics transport models. This mechanism differs from conventional gene regulation mechanisms, which often involve direct modulation of transcription initiation. Our findings point to a genome-wide function of DNA methylation that may account for the reversibility of Rett syndrome in mice. Moreover, our combined theoretical and experimental approach provides a general method for understanding how global gene expression patterns are choreographed.

MeCP2 | Gene regulation | Mathematical modelling | DNA methylation

Introduction

Many eukaryotic chromatin-associated factors modulate transcription by binding to specific sites in gene promoters or enhancers (1, 2). Most transcription factors are thought to modulate the initiation rate of transcription by altering histone-DNA interactions (2, 3) or imposing promoter-proximal obstacles (4). However, transcription can also be affected by processes that occur in the bodies of genes. In particular, DNA methylation, which is widespread in gene bodies, appears to affect progression of RNA polymerase II (RNA Pol II) through densely methylated exons (5). The mechanism is unclear, but methyl-CpG binding proteins (6) may be involved. Since most gene bodies contain methylated CpGs, such proteins may have a global effect on transcription.

One putative global modulator is methyl-CpG binding protein 2 (MeCP2) (7, 8), which is highly expressed in neurons. *MECP2* mutations, including loss-of-function or gene duplication, lead to severe neurological disorders (9, 10). MeCP2 does not behave as a conventional transcription factor with discrete targets, as its binding site occurs on average every ~100 base pairs. Evidence from *in vitro* systems (11, 12) and mouse models (13, 14) suggests that MeCP2 can mediate DNA methylation-dependent transcriptional inhibition. Transcriptional changes in mouse brain when MeCP2 is absent or over-expressed are relatively subtle but widespread (15-17), and the molecular mechanisms underlying these changes are unknown.

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 (18) with computational 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 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 the TBL1/TBLR1 subunits of the NCoR co-repressor complex fails to repress efficiently, suggesting that repression depends upon this interaction.

Results

Global changes in transcription correlate with MeCP2 expression level. We created progenitor cell lines capable of differentiation to a uniform population of human neurons (SI Appendix, Fig. S1A-C) that expressed seven widely different levels of MeCP2, including knock-out (KO), wild-type (WT) and 11-

Significance

Patterns of gene expression are primarily determined by proteins that locally enhance or repress transcription. While many transcription factors target a restricted number of genes, others appear to modulate transcription levels globally. An example is MeCP2, an abundant methylated-DNA binding protein that is mutated in the neurological disorder Rett Syndrome. Despite much research, the molecular mechanism by which MeCP2 regulates gene expression is not fully resolved. Here we integrate quantitative, multi-dimensional experimental analysis and mathematical modelling to indicate that MeCP2 is a novel type of global transcriptional regulator whose binding to DNA creates "slow sites" in gene bodies. Our combined theoretical and experimental approach provides a general method for understanding how gene expression patterns are choreographed.

Reserved for Publication Footnotes

137
138
139
140
141
142
143
144
145
146
147
148
149
150
151
152
153
154
155
156
157
158
159
160
161
162
163
164
165
166
167
168
169
170
171
172
173
174
175
176
177
178
179
180
181
182
183
184
185
186
187
188
189
190
191
192
193
194
195
196
197
198
199
200
201
202
203
204

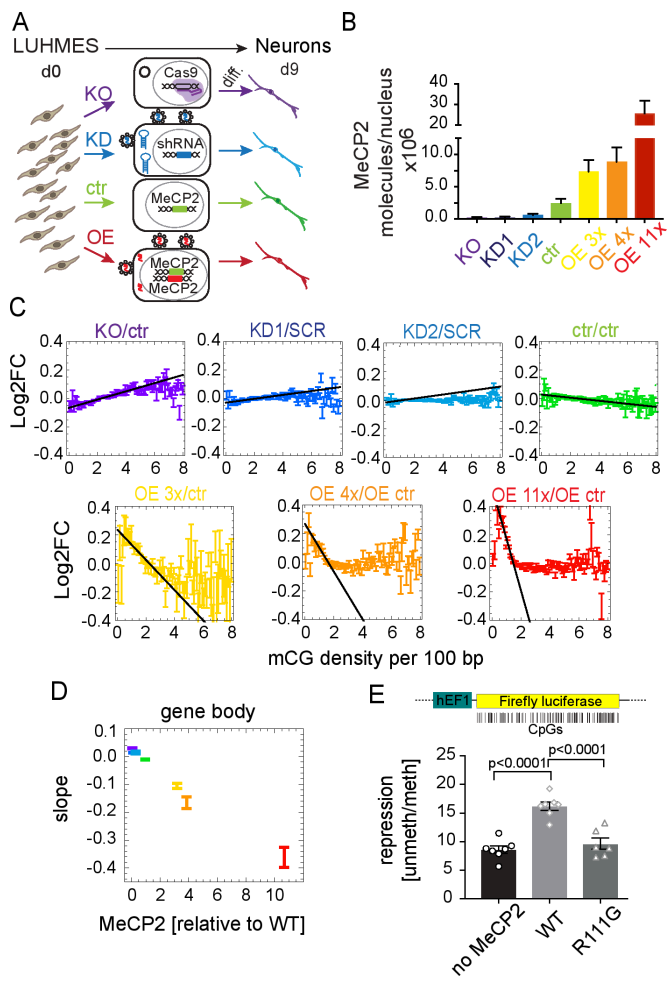


Fig. 1. Gene expression strongly correlates with gene body mCG density and MeCP2 abundance. (A) Experimental design (Methods). (B) Mean number of MeCP2 molecules per nucleus. (C) Log₂ fold change of gene expression (Log₂FC) relative to appropriate controls (ctr – unmodified controls; SCR – scrambled shRNA control, OE ctr – overexpression control) for all seven levels of MeCP2, plotted against gene body mCG density. All Log₂FC values have been shifted so that Log₂FC averaged over all genes is zero. Black line indicates the maximum slope. (D) The maximum slope for gene bodies varies proportionally to MeCP2 abundance. (E) Ratio between luciferase expressions from an unmethylated and gene-body methylated constructs, for three cases: no MeCP2, WT MeCP2, and an MBD mutant R111G that is unable to bind mCG. Points show individual replicates. In all panels, error bars represent \pm SEM.

fold over-expression (11x) (Fig. 1A,B and SI Appendix, Fig. S1D and Table S1). All lines differentiated into neurons with similar kinetics, expressed neuronal markers (SI Appendix, Fig. S1E), and had identical global levels of DNA methylation ($\sim 3.7\%$ of all cytosines were methylated) (SI Appendix, Fig. S2A). Based on the known affinity of MeCP2 for methylated CG (mCG), we expected that the effect of MeCP2 on gene expression would depend on their mCG content. DNA methylation was therefore quantified for all genes in WT neurons using whole-genome bisulfite sequencing (TAB-seq) (SI Appendix, Fig. S2B,C). We calculated total methylation (total mCG, M_{mCG}) as the number of methylated CG dinucleotides, mCG density (ρ_{mCG}) as the number of mCGs per 100 bp, and mCG mean as the percentage of mCG in all CG dinucleotides. To determine the effects of MeCP2 on transcription, we performed RNA-seq on all seven cell lines. We included all expressed protein-coding genes (~ 17000 genes) in our analysis. Most genes responded to MeCP2 but changes

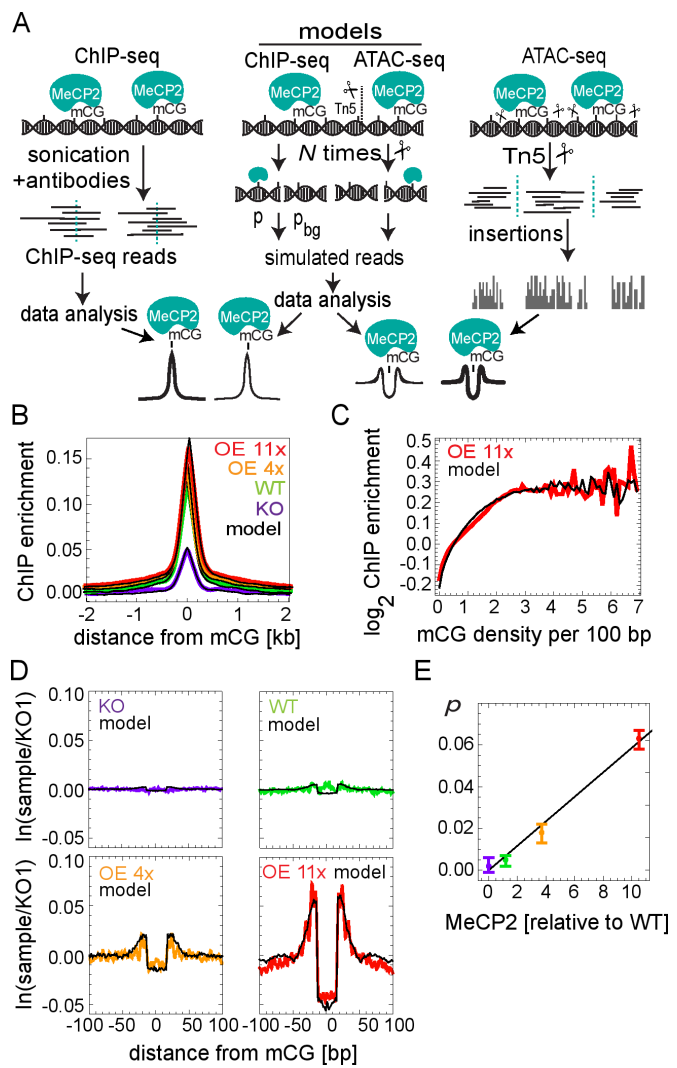


Fig. 2. MeCP2 occupancy on the DNA is proportional to mCG density and MeCP2 level. (A) MeCP2 ChIP- and ATAC-seq experimental procedures and their *in silico* counterparts. and are probabilities of background and MeCP2-bound reads, respectively. Tn5 insertion sites (scissors) occur in exposed DNA regions. (B) ChIP-seq enrichment profiles centered at mCG dinucleotides for different cell lines. Black lines represent *in silico* profiles fitted to the experimental data. (C) MeCP2 ChIP-seq enrichment data in OE 11x/KO (red) as a function of mCG density. (D) Average depletion profiles (logarithm of the ratio between the number of Tn5 insertions in a given cell line and KO1, 2-4 biological replicates) in the ± 100 bp regions surrounding mCG dinucleotides. Black lines represent computer simulations of the model fitted to the data. (E) Predicted fraction of mCGs occupied by MeCP2 versus MeCP2 level obtained from depletion profiles in (D). Error bars represent \pm SEM.

were small, precluding definition of a subset of affected genes (SI Appendix, Fig. S3A). To enhance a possible relationship between expression changes and DNA methylation that otherwise might be obscured by other regulatory mechanisms and statistical noise, genes were binned according to methylation density, considering gene bodies and promoters separately.

The average change in expression versus appropriate controls (Log₂FC) showed a strong relationship to mCG density (ρ_{mCG}) in gene bodies (Fig. 1C). The effect was the strongest for $\rho_{mCG} = 0.8-4.0$ mCG per 100bp which includes the vast majority of genes (SI Appendix, Fig. S3B). The apparent stimulation of expression at very low mCG densities in OE neurons is discussed in SI Appendix. Moreover, the maximum slope of the Log₂FC versus ρ_{mCG} in gene bodies (Fig. 1C, black lines) was strikingly

205
206
207
208
209
210
211
212
213
214
215
216
217
218
219
220
221
222
223
224
225
226
227
228
229
230
231
232
233
234
235
236
237
238
239
240
241
242
243
244
245
246
247
248
249
250
251
252
253
254
255
256
257
258
259
260
261
262
263
264
265
266
267
268
269
270
271
272

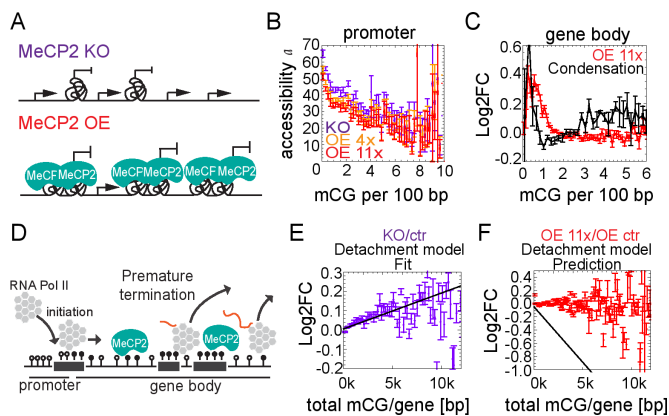


Fig. 3. MeCP2 does not regulate transcription via condensation of chromatin or premature termination. (A) A cartoon of the Condensation model. Tangles represent regions of condensed chromatin that are inaccessible to RNA Pol II. (B) Chromatin accessibility (measured by ATAC-seq) at promoters rapidly decreases with increasing promoter methylation. In contrast, MeCP2 has a minor effect on accessibility (curves for OEs 4x and 11x are slightly lower than for KO). (C) The Condensation model disagrees with Log2FC(OE 11x/KO) obtained from RNA-seq. (D) Schematic representation of the Detachment model. (E) Log2FC (gene expression) for KO/ctr (purple) versus the total number of mCGs per gene. Black lines represent predictions of Detachment model. Error bars represent \pm SEM. (F) As (E) for OE 11x/OE ctr (red).

proportional to MeCP2 levels (Fig. 1D). In contrast, plots of Log2FC versus ρ_{mCG} in promoter regions showed a slope close to zero, indicating minimal dependence on promoter methylation (SI Appendix, Fig. S3C). No clear dependence on MeCP2 level was observed for Log2FC versus total gene body mCG or mCG mean (SI Appendix, Fig. S3D,E). These results indicated that the gene-body mCG density is the strongest predictor of MeCP2-dependent transcriptional changes. This relationship is not affected when data are filtered by significance, gene length or promoter methylation (SI Appendix, Fig. S4A-D). Moreover, the relationship is maintained even when intronic reads are analyzed suggesting that pre-mRNA is affected in the same way as processed RNA (SI Appendix, Fig. S4E). To test for a causal relationship, we transfected cells with two versions (methylated or unmethylated gene body) of a luciferase reporter gene with a methylation-free promoter in the presence of wildtype or the DNA binding mutant MeCP2[R111G] (SI Appendix, Fig. S5A,B). We observed a two-fold repression of methylated versus unmethylated luciferase gene body in the presence of WT MeCP2 compared to either no MeCP2 or mutant MeCP2 (Fig. 1E).

MeCP2 binds predominantly methylated CG genome-wide.

To map the binding of MeCP2 in human neurons, we performed 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* (Fig. 2A). As expected, ChIP enrichment was proportional to the level of MeCP2 in each cell line (SI Appendix, Fig. S6A-C) and showed a strong peak centred at mCGs in MeCP2-positive lines (Fig. 2B) as well as a correlation between MeCP2 enrichment and mCG density (Fig. 2C). Conversely, enrichment was absent at non-methylated CGs (SI Appendix, Fig. S6E).

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 (19) in which transposase Tn5 cuts exposed DNA to reveal DNA accessibility within chromatin (Fig. 2A). In agreement with the ChIP-seq data, ATAC-seq Tn5 insertion profiles (Fig. 2D) showed a graded depletion of insertion sites centered around mCG in WT, OE 4x and OE 11x neurons, whose amplitude was proportional to MeCP2 concentration (Fig. 2E) and therefore represents a “molecular

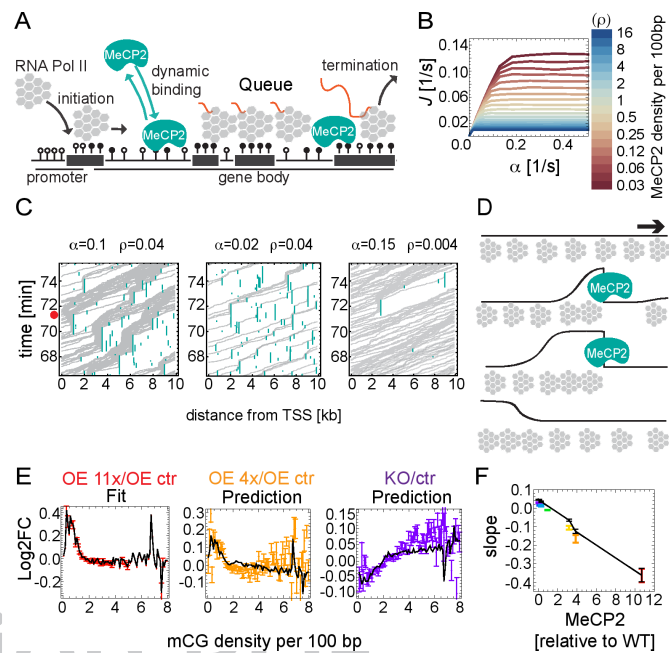


Fig. 4. Mathematical modelling indicates that MeCP2 slows down transcriptional elongation. (A) Schematic representation of the Dynamical obstacles model. (B) Transcription rate predicted by the model, plotted as a function of the initiation rate α , for different mean MeCP2 densities in gene bodies. (C) Space-time plots (kymographs) representing Pol II moving along the gene. Queues of Pol II induced by MeCP2 can reach TSS (red dot) and block initiation if both the initiation rate (α) and the density of MeCP2 (ρ) are sufficiently high (left panel). (D) Schematic representation of Pol II (grey) density shock waves forming behind MeCP2 (blue). Black line is the local density of Pol II. (E) Log2FC (gene expression) versus mCG density in gene bodies obtained in computer simulations of the Dynamical obstacles model (black solid lines) fitted to the OE 11x/OE ctr RNA-seq dataset (red) agrees well with experimental data for OE 4x/OE ctr (orange) and KO/ctr (purple) datasets. Error bars represent \pm SEM. (F) The maximum slope of Log2FC (gene expression) versus mCG density in gene bodies, predicted by the Dynamical obstacles model (black line). Points are experimental slopes from Fig. 1C.

footprint” of MeCP2 binding *in vivo*. The size and amplitude of the footprint agrees well with a computer model of ATAC-seq and MeCP2 binding (Fig. 2D, black lines) and previous *in vitro* data (20, 21), confirming that MeCP2 occupies 11bp of DNA in living cells. No depletion of insertion sites was observed over unmethylated CG (SI Appendix, Fig. S6F). The model revealed that only 6.3% of mCG sites are actually occupied by MeCP2 in OE 11x neurons, falling to less than 1% occupancy in WT (Fig. 2E), perhaps due in part to occlusion by nucleosomes. Excellent agreement between the models and ATAC-seq and ChIP-seq data allows us to predict MeCP2 occupancy from mCG density and MeCP2 level in each cell line (Fig. 2E and SI Appendix, Fig. S6D).

MeCP2 does not regulate transcription via condensation of chromatin or premature termination.

To interpret these results mechanistically, we considered mathematical models based on a commonly accepted paradigm for gene expression (SI Appendix, Fig. S7A) (22). In the first class of models named Condensation models (Fig. 3A), MeCP2 affects the rate of transcription initiation via changes in chromatin structure. The possibility that MeCP2 affects the initiation rate α by binding to promoters was rejected because it would imply a stronger correlation between gene expression and ρ_{mCG} in promoters than in gene bodies, contrary to our observations (SI Appendix, Fig. S3C). MeCP2 could hypothetically affect the fraction f of cells with specific genes in the ON state via some long-distance mechanism involving binding to gene bodies and leading to changes in the degree of chro-

409
410
411
412
413
414
415
416
417
418
419
420
421
422
423
424
425
426
427
428
429
430
431
432
433
434
435
436
437
438
439
440
441
442
443
444
445
446
447
448
449
450
451
452
453
454
455
456
457
458
459
460
461
462
463
464
465
466
467
468
469
470
471
472
473
474
475
476

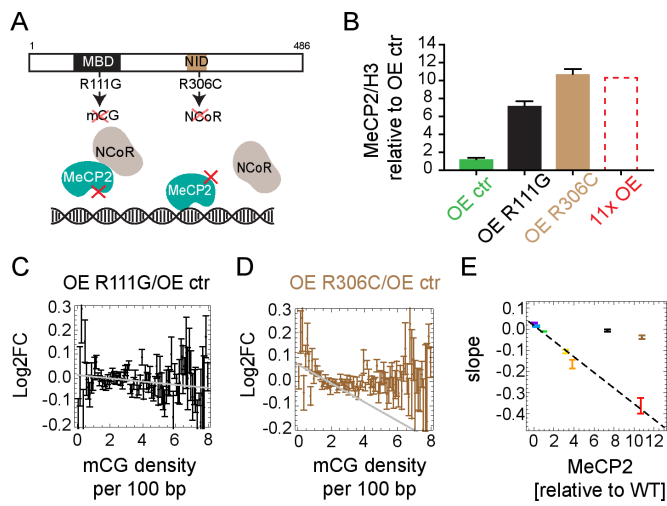


Fig. 5. MeCP2 slows down transcription via a mechanism involving NCoR. (A) Location of two binding domains in MeCP2 that are relevant for the proposed mechanism: methyl-CpG binding domain (MBD) and NCoR-interaction domain (NID). The mutation R111G causes MeCP2 to lose the ability to bind specifically 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). OE 11x is shown for comparison. Values are averaged over three biological replicates and normalised by the level of histone H3. (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. (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 \pm SEM.

matin openness near promoters. However, mapping chromatin accessibility using ATAC-seq showed that while there is a weak correlation between MeCP2 and accessibility (Fig. 3B), it cannot account for the observed Log₂FC in gene expression (Fig. 3C).

We next considered potential effects of MeCP2 on the elongation phase of transcription. The Detachment Model posits that MeCP2 causes transcription to prematurely abort (Fig. 3D). Since the probability of termination increases with each blocking site, under this model the Log₂FC is a function of the total number of methylated CGs (N_{mCG}) in the gene: $\text{Log}_2\text{FC} = -\gamma(M-1)N_{mCG}$, where M is MeCP2 concentration relative to WT, and the parameter γ is proportional to the probability that Pol II aborts transcription when it encounters MeCP2 or an MeCP2-induced chromatin modification. The unknown parameter γ can be obtained by fitting the model to the Log₂FC (KO/WT) data (Fig. 3E). We found that the model failed to reproduce the Log₂FC vs N_{mCG} relationship for the OE 11x cell line (Fig. 3F). The model also fails to correctly predict the observed relationship between Log₂FC and mCG density in gene bodies (SI Appendix, Fig. S7B,C). Therefore, it is unlikely that MeCP2 affects transcription via premature termination.

MeCP2 creates “Dynamical obstacles” that impede transcriptional elongation. Finally, we considered a “Congestion model” whereby Pol II pauses when it encounters MeCP2 itself or an induced, transient structural modification of chromatin (Fig. 4A). The parameters are: the fraction p of mCGs bound by MeCP2, MeCP2 turn-over (unbinding) rate k_u , and (specific to each gene) the length L of the gene, the density ρ_{mCG} of methylated CGs, and the initiation rate α . Fig. 4B shows the transcription rate for OE 11x predicted by the model as a function of α , for different mean MeCP2 densities ($p\rho_{mCG}$). The assumed value of $k_u = 0.04$ s⁻¹ is compatible with the reported *in vivo* residence time of

MeCP2 on chromatin (25-40s (23)). Inspired by non-equilibrium statistical mechanics approaches that have been utilised to model one-dimensional transport (24, 25), we expect a non-equilibrium phase transition from a low-density to a maximal-current (congested) phase as the initiation rate or the density of obstacles increase beyond a critical point. Indeed, all curves in Fig. 4B have a characteristic shape: a linear relationship $J \propto \alpha$ for small α , followed by saturation at high initiation rates. Saturation occurs due to congestion as polymerases queue upstream of obstacles (Movies S1,2). However, even in the non-saturated regime of intermediate α , excluded-volume interactions between polymerases that have been slowed down by an obstacle cause a density shock wave that propagates backwards (Fig. 4C). A small increase in the density of polymerases near the promoter decreases the rate of Pol II binding to the TSS. Thus, even though MeCP2 does not directly affect Pol II initiation, it does so indirectly by shock waves that form behind MeCP2-induced obstacles in gene bodies (Fig. 4D). To test the model against RNA-seq data, we estimated average initiation rates for genes with similar mCG densities by fitting the model to Log₂FC data from one of the cell lines (OE 11x/OE ctr; Fig. 4E left and SI Appendix, Fig. S8F). We then used the model to predict Log₂FC for the remaining 6 cell lines. The model strikingly reproduces the data (Fig. 4E for OE 4x and KO) as well as the slopes of the Log₂FC plots for all seven cell lines (Fig. 4F). A similar behaviour occurs in a modified model in which Pol II slows down (rather than completely stops) on permanent or long-lasting structural modifications of chromatin (SI Appendix, Fig. S8A-E, Movie S3). We conclude that both congestion models are compatible with the experimental data presented in Fig. 1C and D. The models also predict that Log₂FC should decrease with increasing expression (measured as TPMs) in agreement with the data (SI Appendix, Fig. S8G).

MeCP2 binding to both DNA and NCoR are essential to slow down RNA Pol II. To address the question of whether MeCP2 impedes Pol II progression directly by steric interference or indirectly by altering chromatin structure (e.g., by histone deacetylation (26)), we overexpressed mutated forms of MeCP2 in the presence of WT MeCP2. The mutants were either unable to bind methylated DNA (R111G) (27) or unable to recruit the histone deacetylase complex NCoR (R306C) (14, 28) (Figs. 5A and SI Appendix, Fig. S9A). As expected, 7-fold overexpression of MeCP2-R111G caused no mCG-density dependent transcriptional changes (Figs. 5B,C and SI Appendix, Fig. S9B,C). The R306C mutant, on the other hand, was predicted to repress transcription if inhibition is directly due to MeCP2 binding to DNA, but not if inhibition is mediated via the corepressor. In fact, 11-fold overexpression of MeCP2-R306C relative to WT MeCP2 caused only a small perturbation of gene expression, indicating a significant loss of DNA methylation-dependent repression (Figs. 5B,D and SI Appendix, Fig. S9B,C). The weak slope may represent minor direct interference of DNA-bound MeCP2-R306C with transcription. As neither mutant falls on the line defining the linear relationship between gene repression and MeCP2 concentration (Fig. 5E), our findings favour a predominantly indirect mechanism of repression, whereby corepressor recruitment alters the chromatin state to impede transcription.

Concluding remarks
In summary, a close alliance between mathematical modelling and molecular biology has allowed us to discriminate molecular mechanisms underlying the relatively subtle global effects of MeCP2 on global gene expression. The proposed mechanism relies on MeCP2-NCoR interaction that slows down the progression of Pol II during transcription elongation. A candidate mediator of this effect is histone modification, in particular histone deacetylation, as cell transfection assays using methylated reporters demonstrate that repression depends upon histone deacetylase activity (11, 12). According to this scenario, MeCP2

477
478
479
480
481
482
483
484
485
486
487
488
489
490
491
492
493
494
495
496
497
498
499
500
501
502
503
504
505
506
507
508
509
510
511
512
513
514
515
516
517
518
519
520
521
522
523
524
525
526
527
528
529
530
531
532
533
534
535
536
537
538
539
540
541
542
543
544

545 recruitment of the histone deacetylase corepressor NCoR would
 546 restrain transcription, perhaps by causing tighter binding of nucle-
 547 osomes to DNA (26). To explain the dramatic reversibility of Rett
 548 syndrome in animal models (29) we propose that, in the absence
 549 of MeCP2, DNA methylation patterns are unaffected, allowing
 550 the re-expressed wildtype protein to bind within gene bodies and
 551 commence normal modulation of transcriptional elongation. We
 552 suggest that the Congestion model may apply to proteins other
 553 than MeCP2. For example, other chromatin-binding factors that
 554 bind short (and thus abundant) motifs, including other methyl-
 555 binding proteins, may modulate gene expression by a similar
 556 mechanism.

557 Materials and methods

558 **Cell lines.** The procedure for culture and differentiation of the LUHMES cell
 559 line was previously described (18). To create two independent *MECP2* knock-
 560 out lines, we used CRISPR-mediated gene disruption (30). To generate MeCP2
 561 knock-downs, several shRNAs against MeCP2 were designed using Sigma-
 562 Aldrich Mission shRNA online software. Two shRNAs were chosen and cloned
 563 into pLKO.1 vector including scrambled shRNA as a control and lentiviruses
 564 were created (SI Appendix, Table S2). To increase the level of MeCP2 we
 565 created lentiviruses expressing MeCP2 from two alternative promoters in the
 566 pLKO.1 vector: Synapsin and cytomegalovirus (CMV). Calculation of standard
 567 deviation, standard error of mean and *t* tests for qPCR, Western blots,
 568 methylation and total RNA quantification using HPLC were performed using
 569 GraphPad Prism version 7.

568 **Repression assay.** CpG-free vector containing Firefly Luciferase with
 569 CpGs was methylated by M.Sss1 methyltransferase in presence or absence
 570 of SAM. Mouse embryonic fibroblasts were transfected using Lipofectamine
 571 2000 with three plasmids containing: Firefly Luciferase, Renilla Luciferase
 572 and MeCP2. Luciferase activity measurements were performed using Dual
 573 Luciferase assay kit (Promega) according to manufacturer protocol.

573 **Library preparation for Illumina sequencing.** All libraries were se-
 574 quenced as 75- or 100-nucleotide long paired-end reads on HiSeq 2000
 575 and HiSeq 2500 Illumina platforms. Methylation of wildtype LUHMES-derived
 576 neurons at day 9 was obtained by TAB-seq according to the published
 577 protocol (31). RNA-seq library was performed according to manufacturer
 578 protocol for ScriptSeq Complete Gold kit (Human/Mouse/Rat). Total RNA
 579 was isolated from all generated cell lines (SI Appendix, Table S1) at day 9
 580 of differentiation using either the RNeasy Mini kit or the AllPrep DNA/RNA
 581 Mini kit (Qiagen). ATAC-seq in four cell lines (KO, WT, OE 4x and OE 11x, SI
 582 Appendix, Table S1) was performed as in (32).

581 MeCP2 ChIP-seq was performed using LUHMES-derived neurons at day
 582 9 of differentiation with four levels of MeCP2: KO, WT, OE 4x and OE 11x (SI
 583 Appendix, Table S1). Libraries were prepared using the NEBNext Ultra II DNA
 584 library Prep kit (NEB) for both IPs and corresponding inputs.

584 **Data processing of raw reads from Illumina sequencing.** All reads
 585 were quality-controlled, trimmed to remove adapters (Trimmomatic) (33),
 586 and duplicated reads, and mapped to the human hg19 reference genome.
 587 Bismark (34) was used to extract cytosine methylation from TAB-seq. All raw
 588 data were deposited in GEO database (accession number GSE125660).

588 **RNA-seq data analysis.** We used a subset of protein-coding genes with
 589 sufficient methylation coverage (BS-seq; $\geq 80\%$ C detected as methylated,
 590 coverage ≥ 20), and gene bodies 1kb or longer. This resulted in 15382 genes
 591 out of the initial 17764 protein-coding genes (86%). In all plots of Log2FC
 592 of differential gene expression we shifted the Log2FC values so that the
 593 average Log2FC in the range of mCG density $\rho_{mCG} \in [1,6]$ 100bp was zero for
 594 all samples. This was motivated by a difficulty in determining the absolute
 595 levels of expression since we did not quantify total mRNA.

595 **ChIP-seq enrichment profiles.** We first obtain accumulated counts (the
 596 number of reads) c_i^x that overlap with *i*-th basepair to the right ($i > 0$) or
 597 left ($i < 0$) from feature *x* ($x = \text{mCG, mCA, ...}$). We then calculate enrichment
 598 profiles as

$$599 f_i = \frac{\text{Norm}_{m_1}(c^{\text{ChIP},x})[i]}{\text{Norm}_{m_1}(c^{\text{input},x})[i]} - 1,$$

600 where $c_i^{\text{ChIP},x}$ and $c_i^{\text{input},x}$ are accumulated counts from ChIP and input (ge-
 601 nomic) DNA sequencing, respectively, and $\text{Norm}_{m_1}(c)[i]$ normalizes the counts
 602 profiles such that their flanks have values close to one:

$$603 \text{Norm}_{m_1}(c)[i] = \frac{c_i}{(\sum_{j=-301}^{-500} c_j + \sum_{j=301}^{500} c_j)/400}.$$

604 We consider a particular C to be methylated if it is methylated in 100% of
 605 the reads, and the coverage is at least 5. We consider a C to be unmethylated
 606 if it does not show up in any of the ChIP-seq reads as methylated.

607 **Computer model of ChIP-seq.** We assume that MeCP2 occupies methyl-
 608 ated cytosines with probability \mathcal{P} times the probability of binding to a
 609 particular motif. Binding probabilities for different motifs are based on

610 known binding affinities (35) and relative binding strengths (15). To create
 611 simulated ChIP fragments, we assume that if a DNA fragment contains at
 612 least one MeCP2 bound to it, it will be present in the simulated ChIP-seq.
 613 Fragments that do not contain any MeCP2 may still be present in the ChIP-seq
 614 data with probability \mathcal{P}_{bg} which accounts for "background" reads in ChIP-seq
 615 even in the absence of MeCP2. This is similar to previous models of ChIP-seq
 616 (36); even best ChIP-seq libraries can have a significant level of noise (\mathcal{P}_{bg} close
 617 to 1) (37). We also add CG- and length bias, and process simulated reads in
 618 the same way as the experimental ChIP data.

619 For each ChIP-seq data set we fitted the simulated profile (parametrized
 620 by $\mathcal{P}, \mathcal{P}_{bg}$) to the experimental profile. Any $\mathcal{P} \leq 0.1$ gives a good fit (SI
 621 Appendix, Fig. S6D), indicating that $\mathcal{P} \approx 0.1$ is the upper bound on mCG
 622 occupancy in 11x OE. We used best-fit parameters to predict profiles on
 623 features other than mCG (SI Appendix, Fig. S6E).

624 **ATAC-seq footprints.** ATAC-seq was analysed in a similar way to ChIP-
 625 seq, except that we used fragments' endpoints (Tn5 insertion sites) to
 626 generate accumulated counts n_i . We calculated the insertion profiles as

$$627 f_i = \ln \left[\frac{\text{Norm}_{m_2}(n_i^{\text{allIn}})}{\text{Norm}_{m_2}(n_i^{\text{KO}})} \right],$$

628 where n_i^{allIn} and n_i^{KO} are the insertion counts profiles for a given cell line
 629 and KO1, respectively, and Norm_{m_2} normalizes the counts profiles such that
 630 their flanks have values close to one:

$$631 \text{Norm}_{m_2}(n_i) = \frac{n_i}{(\sum_{j=-41}^{-50} n_j + \sum_{j=41}^{50} n_j)/20}.$$

632 **Computer model of ATAC-seq.** We use the same binding model as in the
 633 ChIP-seq simulations. We assume that MeCP2 occupies 11bp (20) and that the
 634 protein is centred on an mC. We simulate the action of the Tn5 transposase by
 635 splitting the sequence into fragments in areas free of MeCP2, and we include
 636 Tn5 sequence bias, and CG- and length bias. The model has three parameters:
 637 the density \mathcal{P} of MeCP2 on mCxx, the average density of insertion (cut) sites
 638 t , and the GC bias b . We process simulated DNA fragments in the same way
 639 as described above for the experimental data. We examined the role of the
 640 parameters on the shape and depth of the simulated footprint of MeCP2 and
 641 concluded that the footprint is not affected as long as the test and control
 642 samples have been processed in a similar way. To extract MeCP2 occupancy
 643 \mathcal{P} from ATAC-seq data, we fitted the model (free parameters \mathcal{P}, t , and a fixed
 644 $b = 6.0$) to experimental footprints for all four cell lines. The relationship is
 645 linear (Fig. 2E), with the best-fit $\mathcal{P} = 0.0058 \times M_{\text{allIn}}/M_{\text{WT}}$.

646 **Chromatin accessibility from ATAC-seq.** For each gene, we calculated
 647 its mean insertion count \bar{n} and selected regions ("insertion peaks") in which
 648 $n_i > 4\bar{n}$. Accessibility was defined as the sum of all insertions in the peaks
 649 divided by the "background" \bar{n} :

$$650 \alpha = \frac{\sum_i n_i}{\bar{n}}.$$

651 **Mathematical models of gene expression.** The condensation model assumes
 652 that the fraction f_i of cells in which gene *i* is actively transcribed depends
 653 on promoter openness α_i (measured by ATAC-seq) which in turn depends on
 654 the level M of MeCP2 and gene methylation ρ_i : $f_i = f_i(M, \rho_i) \propto \alpha_i = \alpha_i(M, \rho_i)$.
 655 The model predicts that $\text{Log2FC}_{X/Y}$ of the ratio of gene expression of cell
 656 line X versus cell line Y should yield the same curve (plus a constant) as the
 657 logarithm of the ratio of accessibilities of X versus Y when plotted as a
 658 function of ρ_{mCG} . Data does not support this model (Fig. 3C). The detachment
 659 model poses that the probability that RNA Pol II successfully terminates is
 660 $P = (1 - \lambda)^n \approx e^{-\lambda n}$, where n is the number of "abort sites" on the gene,
 661 proportional to the number of MeCP2 molecules on the gene, and λ is the
 662 abortion probability. We show that

$$663 \text{Log2FC}_{X/Y} = \text{const} - \gamma \left(\frac{M_X}{M_Y} - 1 \right) n,$$

664 where $\gamma \propto \lambda$ is an unknown parameter identical for all cell lines, and M_X, M_Y
 665 are MeCP2 levels in cell lines X and Y. The model is rejected (Fig. 3F).

666 We consider two mechanisms by which MeCP2 could affect elongation.
 667 To implement the slow sites model we use the totally asymmetric simple
 668 exclusion process (TASEP) with open boundaries (24). A gene is represented
 669 as a chain of L sites. Each site (equivalent to 60bp of the DNA) is either
 670 occupied by a particle (RNA Pol II) or is empty. Particles enter the chain at
 671 site $i = 1$ with rate α (transcription initiation rate), move along the chain
 672 and exit at site $i = L$ with rate $\beta = 1 \text{ sec}^{-1}$. Sites can be "fast" or "slow". Slow
 673 sites represent mCGs affected by the interaction with MeCP2, whereas fast
 674 sites are all other sites (methylated or not). Particles jump with rate $\nu = 1$
 675 sec^{-1} (equivalent to Pol II speed $\approx 60\text{bp/s}$) on fast sites and $\nu_s = 0.05 \text{ sec}^{-1}$
 676 on slow sites. Slow sites are randomly and uniformly distributed with density
 677 $\rho_s = \mathcal{P} \rho_{mCG}$ where \mathcal{P} is the probability that an mCG is occupied by MeCP2. To
 678 relate this model to the mRNA-seq differential expression data we calculate
 679 Log2FC as

681
682
683
684
685
686
687
688
689
690
691
692
693
694
695
696
697
698
699
700
701
702
703
704
705
706
707
708
709
710
711
712
713
714
715
716
717
718
719
720
721
722
723
724
725
726
727
728
729
730
731
732
733
734
735
736
737
738
739
740
741
742
743
744
745
746
747
748

$$\text{Log2FC}_{X/Y} = \log_2 \frac{j(\alpha, \rho_{aX})}{j(\alpha, \rho_{aY})},$$

where $\rho_{aX} = \rho_{mCG} p_X$, $\rho_{aY} = \rho_{mCG} p_Y$ in which p_X, p_Y are MeCP2 occupation probabilities for cell lines X, Y. In the above expression we know all quantities except the initiation rate α which we fit to the OE 11x data.

The dynamical obstacles model is very similar with two exceptions: (i) Pol II always moves with the same speed v (no slow sites) as long as it is not blocked by other polymerases and obstacles, (ii) obstacles bind and unbind dynamically from the methylated sites. We assume that unbinding occurs with rate k_u per obstacle, whereas binding occurs with rate $k_b p$ per unoccupied mCG. Obstacles do not bind if an mCG is already occupied by an obstacle or a polymerase. We assume that obstacles are not restricted to accessible mCGs and that their density on actively transcribed genes may be higher than p obtained from ATAC-seq but still proportional to MeCP2 level. We found that $p = M/M_{OB1x}$ reproduces Log2FC data for all cell lines.

Additional details for Materials and Methods are provided in SI Appendix.

1. Lin CY, et al. (2012) Transcriptional Amplification in Tumor Cells with Elevated c-Myc. *Cell* 151(1):56–67.
2. Berry S, Dean C, Howard M (2017) Slow Chromatin Dynamics Allow Polycomb Target Genes to Filter Fluctuations in Transcription Factor Activity. *Cell Systems* 4(4):445–457.e8.
3. Ouararhni K, et al. (2006) The histone variant mH2A1.1 interferes with transcription by down-regulating PARP-1 enzymatic activity. - PubMed - NCBI. *Genes Dev* 20(23):3324–3336.
4. Hao N, Palmer AC, Dodd IB, Shearwin KE (2017) Directing traffic on DNA—How transcription factors relieve or induce transcriptional interference. *Transcription* 8(2):120–125.
5. Jonkers I, Lis JT (2015) Getting up to speed with transcription elongation by RNA polymerase II. *Nat Rev Mol Cell Biol* 16(3):167–177.
6. Hendrich B, Bird A (1998) Identification and characterization of a family of mammalian methyl-CpG binding proteins. *Molecular and Cellular Biology* 18(11):6538–6547.
7. Lewis JD, et al. (1992) Purification, sequence, and cellular localization of a novel chromosomal protein that binds to Methylated DNA. *Cell* 69(6):905–914.
8. Skene PJ, et al. (2010) Neuronal MeCP2 is expressed at near histone-octamer levels and globally alters the chromatin state. *Molecular Cell* 37(4):457–468.
9. Amir RE, et al. (1999) Rett syndrome is caused by mutations in X-linked MECP2, encoding methyl-CpG-binding protein 2. *Nat Genet* 23(2):185–188.
10. Ramocki MB, Tavayev YI, Peters SU (2010) The MECP2 Duplication Syndrome. *American journal of medical genetics Part A* 152A(5):1079–1088.
11. Nan X, Campoy FJ, Bird A (1997) MeCP2 is a transcriptional repressor with abundant binding sites in genomic chromatin. *Cell* 88(4):471–481.
12. Nan X, et al. (1998) Transcriptional repression by the methyl-CpG-binding protein MeCP2 involves a histone deacetylase complex. - PubMed - NCBI. *Nature* 393(6683):386–389.
13. Guy J, Hendrich B, Holmes M, Martin JE, Bird A (2001) A mouse Mecp2-null mutation causes neurological symptoms that mimic Rett syndrome. *Nat Genet* 27(3):322–326.
14. Lyst MJ, et al. (2013) Rett syndrome mutations abolish the interaction of MeCP2 with the NCoR/SMRT co-repressor. *Nat Neurosci* 16(7):898–902.
15. Lagger S, et al. (2017) MeCP2 recognizes cytosine methylated tri-nucleotide and di-nucleotide sequences to tune transcription in the mammalian brain. *PLoS Genet* 13(5):e1006793.
16. Kinde B, Wu DY, Greenberg ME, Gabel HW (2016) DNA methylation in the gene body influences MeCP2-mediated gene repression. *Proceedings of the National Academy of Sciences* 113(52):15114–15119.
17. Gabel HW, et al. (2015) Disruption of DNA-methylation-dependent long gene repression in Rett syndrome. *Nature*:1–21.
18. Scholz D, et al. (2011) Rapid, complete and large-scale generation of post-mitotic neurons from the human LUHMES cell line. *J Neurochem* 119(5):957–971.
19. Buenrostro JD, Giresi PG, Zaba LC, Chang HY, Greenleaf WJ (2013) Transposition of native chromatin for fast and sensitive epigenomic profiling of open chromatin, DNA-binding

Acknowledgments

We thank Tanja Waldmann for introducing us to the LUHMES cell line, Beatrice Alexander-Howden for technical support, David Kelly for microscopy assistance, Martin Waterfall for help with FACS and Jim Selfridge for help in 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. AB is a member of the Simons Initiative for the Developing Brain. 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** J.C.W. and R.S. designed experiments, made and characterised cell lines and interpreted data; J.C.W. performed the NGS experiments; S.W. and K.C. performed NGS data preparation; B.R. performed HPLC analysis; M.Y. performed initial reaction for TAB-seq; P.G. and B.W. developed the “Congestion model”. B.W. conceived the mathematical modelling, analysed NGS data and interpreted data. A.B. conceived the study and, together with J.C.W., R.S. and B.W., wrote and edited manuscript.

- proteins and nucleosome position. *Nat Meth* 10(12):1213–1218.
20. Nan X, Meehan RR, Bird A (1993) Dissection of the methyl-CpG binding domain from the chromosomal protein MeCP2. *Nucleic Acids Research* 21(21):4886–4892.
21. Nikitina T, et al. (2007) MeCP2-chromatin interactions include the formation of chromosome-like structures and are altered in mutations causing Rett syndrome. - PubMed - NCBI. *J Biol Chem* 282(38):28237–28245.
22. Shahrezaei V, Ollivier JF, Swain PS (2008) Colored extrinsic fluctuations and stochastic gene expression. *Molecular Systems Biology* 4:196.
23. Klose RJ, et al. (2005) DNA Binding Selectivity of MeCP2 Due to a Requirement for A/T Sequences Adjacent to Methyl-CpG. *Molecular Cell* 19(5):667–678.
24. Blythe RA, Evans MR (2007) Nonequilibrium steady states of matrix-product form: a solver's guide. *J Phys A: Math Theor* 40(46):R333–R441.
25. Derrida B (1998) An exactly soluble non-equilibrium system: The asymmetric simple exclusion process. *Physics Reports* 301(1–3):65–83.
26. Zentner GE, Henikoff S (2013) Regulation of nucleosome dynamics by histone modifications. *Nature Publishing Group* 20(3):259–266.
27. Kudo S, et al. (2003) Heterogeneity in residual function of MeCP2 carrying missense mutations in the methyl CpG binding domain. *J Med Genet* 40(7):487–493.
28. Kruusvee V, et al. (2017) Structure of the MeCP2–TBLR1 complex reveals a molecular basis for Rett syndrome and related disorders. *Proc Natl Acad Sci USA* 114(16):E3243–E3250.
29. Guy J, Gan J, Selfridge J, Cobb S, Bird A (2007) Reversal of Neurological Defects in a Mouse Model of Rett Syndrome. *Science* 315(5815):1143–1147.
30. Shah RR, et al. (2016) Efficient and versatile CRISPR engineering of human neurons in culture to model neurological disorders. *Wellcome Open Res* 1:13.
31. Yu M, et al. (2012) Base-resolution analysis of 5-hydroxymethylcytosine in the mammalian genome. *Cell* 149(6):1368–1380.
32. Buenrostro JD, Wu B, Chang HY, Greenleaf WJ (2015) *ATAC-seq: A Method for Assaying Chromatin Accessibility Genome-Wide* (John Wiley & Sons, Inc., Hoboken, NJ, USA).
33. Bolger AM, Lohse M, Usadel B (2014) Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics* 30(15):2114–2120.
34. Krueger F, Andrews SR (2011) Bismark: a flexible aligner and methylation caller for Bisulfite-Seq applications. - PubMed - NCBI. *Bioinformatics* 27(11):1571–1572.
35. Sperlazza MJ, Bilinovich SM, Sinanan LM, Javier FR, Williams DC Jr (2017) Structural Basis of MeCP2 Distribution on Non-CpG Methylated and Hydroxymethylated DNA. *Journal of Molecular Biology* 429(10):1581–1594.
36. Xu H, et al. (2010) A signal-noise model for significance analysis of ChIP-seq with negative control. - PubMed - NCBI. *Bioinformatics* 26(9):1199–1204.
37. Liang K, Keles S (2012) Normalization of ChIP-seq data with control. *BMC Bioinformatics* 13(1):199.

749
750
751
752
753
754
755
756
757
758
759
760
761
762
763
764
765
766
767
768
769
770
771
772
773
774
775
776
777
778
779
780
781
782
783
784
785
786
787
788
789
790
791
792
793
794
795
796
797
798
799
800
801
802
803
804
805
806
807
808
809
810
811
812
813
814
815
816