Quantitative analysis questions the role of MeCP2 in alternative splicing

Kashyap Chhatbar[1,2], Justyna Cholewa-Waclaw[2], Ruth Shah[2], Adrian Bird[2], and Guido Sanguinetti[1,3,✉]

[1]School of Informatics, University of Edinburgh, Edinburgh, United Kingdom
[2]The Wellcome Centre for Cell Biology, University of Edinburgh, Edinburgh, United Kingdom
[3]International School for Advanced Studies (SISSA), Trieste, Italy

MeCP2 is an abundant protein in mature nerve cells, where it binds to DNA sequences containing methylated cytosine. Mutations in the *MECP2* gene cause the severe neurological disorder Rett syndrome (RTT), provoking intensive study of the underlying molecular mechanisms. Multiple functions have been proposed, one of which involves a regulatory role in splicing. Here we leverage the recent availability of high-quality transcriptomic data sets to probe quantitatively the potential influence of MeCP2 on alternative splicing. Using a variety of machine learning approaches that can capture both linear and non-linear associations, we show that widely different levels of MeCP2 have a minimal effect on alternative splicing in three different systems. Alternative splicing was also apparently indifferent to developmental changes in DNA methylation levels. Our results suggest that regulation of splicing is not a major function of MeCP2. They also highlight the importance of multi-variate quantitative analyses in the formulation of biological hypotheses.

MeCP2 | DNA methylation | Machine learning | Splicing
Correspondence: ***gsanguin@sissa.it***

Introduction

MeCP2 (methyl CpG binding protein 2) is an important mediator of epigenetic regulation in mammalian cells, and particularly in neurons(1–3). Mutations that lead to depletion or overexpression of the *MECP2* gene are associated with severe neurological diseases, most notably Rett syndrome, a devastating autism-spectrum disease affecting approximately 1 in 10,000 female births(4). MeCP2 is a chromatin protein that binds to modified cytosine residues, primarily mCG(5), the most abundant modification in mammalian genomes, but also mCA, particularly mCAC(6), which is a feature of neuronal cells(7, 8). This association with chromatin enables MeCP2 to interpret epigenetic signals to modulate gene expression events. By far the best documented role of MeCP2 is repression of transcription: several large-scale studies have confirmed this role, and proposed mechanistic models that can account for the effect of MeCP2 on transcription(9–11). Other roles for MeCP2 have also been hypothesised, however, including chromatin compaction, micro-RNA processing, and regulation of alternative splicing (reviewed in reference(12)). These hypotheses highlight alternative roles of MeCP2 in orchestrating gene expression at transcriptional and post-transcriptional levels, but they have varying degrees of support from data.

In this paper, we focus on the possibility that MeCP2 is a regulator of alternative splicing. This was first proposed after a physical interaction with the splicing factor YB1 was observed(13) in 2005. Subsequent studies provided additional support; most recently via a modest but significant association between changes in DNA methylation and intron retention in the transition between promyelocytes and granulocytes(14). A role of MeCP2 in splicing is plausible from a mechanistic perspective(9–11) if binding of the protein serves as a "brake" for polymerase, leading to a reduced transcription rate, which in turn can alter splicing preferences(15). These studies relied on correlative, univariate analyses of data from experiments which were not designed to specifically test the association of MeCP2 with splicing. This renders the conclusions vulnerable to confounding effects, which could explain both changes in splicing and DNA methylation/MeCP2 occupancy. Although a measure of significance of the association (assuming no confounding factors) was deduced, the effect size was not quantified. As a result, questions regarding the extent to which variation in splicing can be attributed to changes in DNA methylation or MeCP2 levels remain open.

Here we revisit the evidence for this phenomenon, leveraging recent, high-quality data sets which track transcriptomic changes in experiments specifically designed to reflect changes in MeCP2 activity. These new data-sets offer an unprecedented opportunity to tackle this question statistically, providing a quantitative estimate of MeCP2's effect while carefully controlling for confounding effects.

Results

It has been suggested that DNA methylation regulates alternative splicing through different mechanisms(16). We evaluated the relationship from two perspectives: either dependent on MeCP2 or independent of MeCP2. Considering the former first, we sought transcriptomic data sets from cells or tissues with different controlled levels of MeCP2 while keeping DNA methylation constant. Several comprehensive data sets have recently been collected in situations that closely mimic this idealised scenario (Table 1). In particular, Cholewa-Waclaw et al.(9) captured transcriptomic data sets in Lund Human Mesencephalic (LUHMES)-derived human dopaminergic neurons that expressed widely different levels of MeCP2, while total levels of DNA methylation remained constant. To investigate the latter, we examined the *Dnmt1/3a/3b* triple-knockout mouse embryonic

PREPRINT

stem cells (*DNMT-TKO* mESCs) that lack DNA methyltransferase activity(17, 18). Secondly, Stroud et al.(19) profiled the transcriptomes of differentiating mouse neurons in the first few weeks after birth, during which time MeCP2 levels increase before reaching a plateau(1). Concurrently, the level of genomic mCA, a major determinant of MeCP2 binding, also increases in the early postnatal brain(5, 19, 20). These data sets allow us to probe the relationship between DNA methylation and alternative splicing in a more quantitative manner. Similarly, we re-analysed the data sets from promyelocytes and granulocytes which first highlighted the potential link between DNA methylation and alternative splicing.

Differential splicing analysis questions the function of MeCP2 and DNA methylation as a global regulator of alternative splicing. To explore the effects of MeCP2 on splicing and gene expression, we initially performed differential splicing and differential gene expression analyses for a number of data sets which investigated changes in MeCP2 levels (Table 1). Varying the abundance of MeCP2 in cultured human neurons over a wide range: absence (*KO*), 1x (*WT*), 4x and 11x wildtype (*OE*) levels is accompanied by many changes in gene expression relative to wildtype ($KO = 178$; $4xOE = 2548$; $11xOE = 2348$ at a significance threshold of p-adjusted value < 0.05), but relatively few differential splicing events ($KO = 101$; $4xOE = 120$; $11xOE = 73$ at a significance threshold of Bayes factor ≥ 3) (Figure S1A,B). This suggests that changing the levels of MeCP2 results in sizable changes in transcription but negligible changes in splicing.

In *WT* mouse neurons, where levels of DNA methylation and MeCP2 increase during development, we found that the number of differential splicing events (489, 408) is less compared to the number of differential gene expression events (1122, 538) when 8 weeks old and 1 week old Parvalbumin (*Pv*) and Vasoactive intestinal peptide (*Vip*) neurons are compared (Figure S1E,F). Comparisons between 8 weeks old and 3 weeks old *Pv* and *Vip* neurons showed more differential splicing events (60, 527) than differentially regulated genes (4, 118) (Figure S1E,F). Based on this analysis, we could not establish whether MeCP2 and DNA methylation predominantly perturbs transcription or alternative splicing in developing neurons. Because of the increase in CA methylation(19, 20) and the dynamic nature of alternative splicing(21) in brain development, it is necessary to investigate the role of DNA methylation independent of developmental effects and MeCP2. Therefore, we performed similar analysis for *DNMT-TKO* mESCs in comparison to *WT* mESCs. In the absence of DNA methylation, many genes are differentially regulated (7342), but there are relatively few differential splicing events (134) (Figure S1C,D). Based on this analysis, DNA methylation predominantly perturbs transcription, with a much smaller effect on splicing. In a separate data set for *DNMT-TKO* mESCs(17), the number of differentially regulated genes (102) was only marginally higher than number of differential splicing events (71) (Figure S1G,H). Differential splicing analysis was not possible on the RNA-seq data set of Promyelocytes and Granulocytes(14, 22) due to the absence of replicates. The lower number of differential splicing events compared to differentially expressed genes is consistent at different significance thresholds (Figure S1) in all comparisons except in developing neurons. By performing classical differential gene expression and differential splicing analysis, we find that MeCP2 primarily functions as a regulator of gene expression and has minimal effect on splicing. Whether the primary role of DNA methylation and neuronal development is in influencing gene expression or alternative splicing, is uncertain. To test this in a more quantitative manner, we used machine learning techniques in our subsequent analysis.

Sequence features are highly predictive of splicing ratios. Machine learning approaches have previously shown that genomic sequence features are highly predictive of exon inclusion/exclusion ratios (splicing ratios) and the splicing code assembled using these features helped characterise alternative splicing on a genome-wide scale(23, 24). The splicing ratios were calculated using several independent methods, and our analysis indicated a strong correlation between results obtained with different methods (Figure S2A,S2E). Methods that directly model splicing ratios such as Miso(25) and BRIE(26) showed particularly strong agreement (Figure S2A). To quantify the explanatory power of sequence features in predicting splicing ratios in the current data sets, we used two independent multivariate regression models: Linear Regression and Random Forest Regression, the latter being a flexible non-parametric approach. We first validated that the relationship between the splicing code and splicing ratios(23, 24) could be confirmed for all the data sets described in Table 1. We find that both models explain 50%-65% of the variation (Figure 1A,C,D,E left panels). This effect is quantified by the metric fraction of variation explained (FVE) which is analogous to R^2. When using splicing ratios from MISO, we still retain 35%-50% of the explanatory power of sequence-derived features (Figure 1A,C,D,E right panels). We note that predictive models achieved lower accuracy in *Vip* neurons and *Pv* neurons during early mouse brain development (Figure 1B), possibly due to the highly dynamic nature of splicing in early development(21). We confirm that splicing ratios can be accurately regressed against sequence-derived features and this analysis provided positive controls for further analysis.

MeCP2 changes account for a negligible fraction of splicing variation. We then probed whether changes in inclusion/exclusion ratios between different cell types could be explained by changing levels of MeCP2. We refer to differential exon inclusion/exclusion between different cell types as splicing variation (not to be confused with splicing ratios, which are the ratio of inclusive/exclusive isoform abundance within the same sample). Since global changes in transcription correlate with MeCP2 levels(9) and sequence-derived features can predict splicing ratios (Figure 1)(23, 24), our hypothesis is that any role of MeCP2 in regulating splicing should be reflected in improved predictions of differential exon inclusion/exclusion. To allow for the possibility that

PREPRINT

MeCP2 regulation of splicing might be dependent on the local sequence context, we used regression models with both MeCP2 levels and sequence features as input variables.

When we quantified the explanatory power of contrasting MeCP2 levels by regressing splicing variation against sequence features and MeCP2 binding, we could only explain 0-3% of the splicing variation regardless of the quantification method or model (Figure 2 left panel). In some instances, the model fit is extremely poor, as evidenced by negative levels of splicing variation explained (Figure 2 right panel). Therefore, multiple levels of MeCP2 and sequence features cannot predict splicing variation, suggesting that MeCP2 does not regulate splicing, even when sequence context is taken into account.

DNA methylation changes account for a negligible fraction of splicing variation. MeCP2 binding in neurons is strongly associated with DNA methylation in a CG and CA context(6). We asked whether changes in context specific DNA methylation in developing brain, the primary determinant of MeCP2 binding, affects splicing variation. To do so, we calculated the changes in DNA methylation at alternative exons and surrounding regions for both dinucleotides CG and CA (Figure 1 in S1 Appendix) in developing mouse neurons(19). In this instance, splicing variation is the difference in splicing ratio across *Pv* and *Vip* neurons at different ages. We quantified the explanatory power of methylation features by regressing splicing variation against features derived from DNA methylation, as well as sequence features. The fraction of variance explained by the models is 0-10% (Figure 3A left panel). The analysis fails to capture any relationship (linear/non-linear) when splicing variation is regressed against DNA methylation alone (Figure S3A). Notice, however, that the inclusion of sequence features accounts for context-specific effects of DNA methylation, in a scenario where MeCP2 is essential for normal function of nerve cells. Therefore, context specific DNA methylation and sequence features in developing neurons are only at best modest predictors of splicing variation.

We then specifically set out to test whether changes in DNA methylation and their sequence context can predict splicing variation. To investigate the relationship between DNA methylation and alternative splicing independent of MeCP2 and developmental effects, we chose *DNMT-TKO* mESCs which lack DNA methylation(17, 18). Here, splicing variation is the difference in splicing ratio between *DNMT-TKO* mESCs and *WT* mESCs. We quantified the explanatory power of methylation features by regressing splicing variation against DNA methylation and associated DNA sequence features. Both regression models are able to account for a minimal fraction (0-10%) of the observed variance in splicing ratios (Figure 3C left panel). In a separate *DNMT-TKO* mESCs data set, the accuracy was lower (0-5%) (Figure 3B left panel). Therefore, DNA methylation when considered in a scenario independent of MeCP2 and developmental effects cannot predict splicing variation.

The promyelocytes-granulocytes data set is the one where methylation levels along with sequence features does appear to explain a small fraction of variance in splicing variation (5-10%) (Figure 3C). However, even in this case, the fraction of variance explained remains modest when compared to the explanatory power of sequence features in predicting splicing ratios (65% at best and 25% at worst, Figure 1 left panels). It is important to stress that both regression models perform worse when DNA methyation features alone are used to predict splicing variation (Figure S3B), rendering the initial observation moot(14). In addition to the absence of replicates, as noted above, a further reason for exercising caution is that this study did not explicitly control for changes in MeCP2 abundance.

Discussion

MeCP2 is a major interpreter of the cell's epigenetic state whose effects on transcription are extensively categorised and recognised as consistent in multiple systems(2, 6, 9–11). In addition to its effects on gene expression, multiple other potential functions for MeCP2 have been proposed. However, its high abundance in neurons and pervasive presence in the chromatin environment create difficulties in attributing functions to MeCP2 based purely on observational studies. One often cited function of DNA methylation and its cognate interpreter, MeCP2, is regulation of alternative splicing. Here we have used multivariate analysis of recent data sets to examine this hypothesis in systems where either DNA methylation levels or MeCP2 levels are varied either alone or in combination. The results reveal that MeCP2's role as a global regulator of RNA splicing is minimal and therefore of questionable biological significance.

While statistically our conclusion is robust, it must be qualified in several ways. First of all, we focused on detecting a global role in splicing for MeCP2. The fact that MeCP2 variation explains a negligible fraction of global splicing variation does not imply that MeCP2 plays no role in regulating splicing for specific transcripts. Our analysis would not detect a role of this kind, which would require more targeted experiments to characterise and validate any affected genes. A second important reservation is that the small splicing effects that we did detect are not necessarily unimportant. Diseases associated with MECP2 mutations are long-term pathologies, indicating that cells can survive for years in the presence of MeCP2 aberrations. Therefore, it is likely that the molecular effects of MeCP2 will be subtle, a point well-made previously(2, 6, 9–11). Thirdly, it should be noted that our best-controlled data sets and most of our analyses apply to neuronal cells where MeCP2 is both abundant and strongly implicated in neuronal cell physiology. While it is appropriate that questions relating to MeCP2 function should focus on neuronal data sets, this might overlook effects in other cell types. Indeed, we did detect a modest statistical association between combined DNA methylation and DNA sequence features in the promyelocyte-granulocyte transition, although there was no relationship between DNA methylation alone and alternative splicing in this system. MeCP2 was initially implicated in regulation of alternative splicing in this myeloid lineage, but genetic depletion of MeCP2 specif-

PREPRINT

ically in non-brain tissues has no obvious phenotypic consequences, indicating that the protein is dispensable outside the central nervous system(27).

On a more general note, our analysis sounds a note of caution regarding the risks of over-interpretation of experiments that rely on high-throughput data. While such experiments are key in expanding our understanding of molecular physiology, univariate analyses can mask confounding factors. Hence, we argue that biological hypotheses arising from observational studies should always be further validated by carefully designed experiments and supported by multivariate statistical analyses.

ACKNOWLEDGEMENTS

We are grateful to Gabriele Schweikert and Duncan Sproul for their valuable input. We also thank Shaun Webb for bioinformatics and technical assistance.

References

1. Jacky Guy, Hélène Cheval, Jim Selfridge, and Adrian Bird. The Role of MeCP2 in the Brain. *Annual Review of Cell and Developmental Biology*, 27(1):631–652, 2011. doi: 10.1146/annurev-cellbio-092910-154121.
2. Rebekah Tillotson and Adrian Bird. The Molecular Basis of MeCP2 Function in the Brain. *Journal of Molecular Biology*, October 2019. ISSN 0022-2836. doi: 10.1016/j.jmb.2019.10.004.
3. Laura A Lavery and Huda Y Zoghbi. The distinct methylation landscape of maturing neurons and its role in Rett syndrome pathogenesis. *Current Opinion in Neurobiology*, 59:180–188, December 2019. ISSN 0959-4388. doi: 10.1016/j.conb.2019.08.001.
4. Ruthie E. Amir, Ignatia B. Van den Veyver, Mimi Wan, Charles Q. Tran, Uta Francke, and Huda Y. Zoghbi. Rett syndrome is caused by mutations in X-linked MECP2 , encoding methyl-CpG-binding protein 2. *Nature Genetics*, 23(2):185–188, October 1999. ISSN 1546-1718. doi: 10.1038/13810.
5. Brian Hendrich and Adrian Bird. Identification and Characterization of a Family of Mammalian Methyl-CpG Binding Proteins. *Molecular and Cellular Biology*, 18(11):6538–6547, November 1998. ISSN 0270-7306.
6. Sabine Lagger, John C. Connelly, Gabriele Schweikert, Shaun Webb, Jim Selfridge, Bernard H. Ramsahoye, Miao Yu, Chuan He, Guido Sanguinetti, Lawrence C. Sowers, Malcolm D. Walkinshaw, and Adrian Bird. MeCP2 recognizes cytosine methylated tri-nucleotide and di-nucleotide sequences to tune transcription in the mammalian brain. *PLOS Genetics*, 13(5):e1006793, May 2017. ISSN 1553-7404. doi: 10.1371/journal.pgen.1006793.
7. Ryan Lister, Eran A. Mukamel, Joseph R. Nery, Mark Urich, Clare A. Puddifoot, Nicholas D. Johnson, Jacinta Lucero, Yun Huang, Andrew J. Dwork, Matthew D. Schultz, Miao Yu, Julian Tonti-Filippini, Holger Heyn, Shijun Hu, Joseph C. Wu, Anjana Rao, Manel Esteller, Chuan He, Fatemeh G. Haghighi, Terrence J. Sejnowski, M. Margarita Behrens, and Joseph R. Ecker. Global Epigenomic Reconfiguration During Mammalian Brain Development. *Science*, 341(6146):1237905, August 2013. ISSN 0036-8075, 1095-9203. doi: 10.1126/science.1237905.
8. Katherine E. Varley, Jason Gertz, Kevin M. Bowling, Stephanie L. Parker, Timothy E. Reddy, Florencia Pauli-Behn, Marie K. Cross, Brian A. Williams, John A. Stamatoyannopoulos, Gregory E. Crawford, Devin M. Absher, Barbara J. Wold, and Richard M. Myers. Dynamic DNA methylation across diverse human cell lines and tissues. *Genome Research*, 23(3):555–567, January 2013. ISSN 1088-9051, 1549-5469. doi: 10.1101/gr.147942.112. Company: Cold Spring Harbor Laboratory Press Distributor: Cold Spring Harbor Laboratory Press Institution: Cold Spring Harbor Laboratory Press Label: Cold Spring Harbor Laboratory Press Publisher: Cold Spring Harbor Lab.
9. Justyna Cholewa-Waclaw, Ruth Shah, Shaun Webb, Kashyap Chhatbar, Bernard Ramsahoye, Oliver Pusch, Miao Yu, Philip Greulich, Bartlomiej Waclaw, and Adrian P. Bird. Quantitative modelling predicts the impact of DNA methylation on RNA polymerase II traffic. *Proceedings of the National Academy of Sciences*, page 201903549, July 2019. ISSN 0027-8424, 1091-6490. doi: 10.1073/pnas.1903549116.
10. Harrison W. Gabel, Benyam Kinde, Hume Stroud, Caitlin S. Gilbert, David A. Harmin, Nathaniel R. Kastan, Martin Hemberg, Daniel H. Ebert, and Michael E. Greenberg. Disruption of DNA-methylation-dependent long gene repression in Rett syndrome. *Nature*, 522 (7554):89–93, June 2015. ISSN 0028-0836, 1476-4687. doi: 10.1038/nature14319.
11. Benyam Kinde, Dennis Y. Wu, Michael E. Greenberg, and Harrison W. Gabel. DNA methylation in the gene body influences MeCP2-mediated gene repression. *Proceedings of the National Academy of Sciences*, 113(52):15114–15119, December 2016. ISSN 0027-8424, 1091-6490. doi: 10.1073/pnas.1618737114.
12. Matthew J. Lyst and Adrian Bird. Rett syndrome: a complex disorder with simple roots. *Nature Reviews Genetics*, 16(5):261–275, May 2015. ISSN 1471-0064. doi: 10.1038/nrg3897.
13. Juan I. Young, Eugene P. Hong, John C. Castle, Juan Crespo-Barreto, Aaron B. Bowman, Matthew F. Rose, Dongcheul Kang, Ron Richman, Jason M. Johnson, Susan Berget, and Huda Y. Zoghbi. Regulation of RNA splicing by the methylation-dependent transcriptional repressor methyl-CpG binding protein 2. *Proceedings of the National Academy of Sciences*, 102(49):17551–17558, December 2005. ISSN 0027-8424, 1091-6490. doi: 10.1073/pnas.0507856102.
14. Justin J.-L. Wong, Dadi Gao, Trung V. Nguyen, Chau-To Kwok, Michelle van Geldermalsen, Rob Middleton, Natalia Pinello, Annora Thoeng, Rajini Nagarajah, Jeff Holst, William Ritchie, and John E. J. Rasko. Intron retention is regulated by altered MeCP2-mediated splicing factor recruitment. *Nature Communications*, 8:15134, May 2017. doi: 10.1038/ncomms15134.
15. Vahid Aslanzadeh, Yuanhua Huang, Guido Sanguinetti, and Jean D. Beggs. Transcription rate strongly affects splicing fidelity and cotranscriptionality in budding yeast. *Genome Research*, 28(2):203–213, 2018. ISSN 1549-5469. doi: 10.1101/gr.225615.117.
16. Galit Lev Maor, Ahuvi Yearim, and Gil Ast. The alternative role of DNA methylation in splicing regulation. *Trends in Genetics*, 31(5):274–280, May 2015. ISSN 01689525. doi: 10.1016/j.tig.2015.03.002.
17. Ahuvi Yearim, Sahar Gelfman, Ronna Shayevitch, Shai Melcer, Ohad Glaich, Jan-Philipp Mallm, Malka Nissim-Rafinia, Ayelet-Hashahar S. Cohen, Karsten Rippe, Eran Meshorer, and Gil Ast. HP1 Is Involved in Regulating the Global Impact of DNA Methylation on Alternative Splicing. *Cell Reports*, 10(7):1122–1134, February 2015. ISSN 2211-1247. doi: 10.1016/j.celrep.2015.01.038.
18. Silvia Domcke, Anaïs Flore Bardet, Paul Adrian Ginno, Dominik Hartl, Lukas Burger, and Dirk Schübeler. Competition between DNA methylation and transcription factors determines binding of NRF1. *Nature*, 528(7583):575–579, December 2015. ISSN 1476-4687. doi: 10.1038/nature16462.
19. Hume Stroud, Susan C. Su, Sinisa Hrvatin, Alexander W. Greben, William Renthal, Lisa D. Boxer, M. Aurel Nagy, Daniel R. Hochbaum, Benyam Kinde, Harrison W. Gabel, and Michael E. Greenberg. Early-Life Gene Expression in Neurons Modulates Lasting Epigenetic States. *Cell*, 171(5):1151–1164.e16, November 2017. ISSN 0092-8674. doi: 10.1016/j.cell.2017.09.047.
20. Yupeng He and Joseph R. Ecker. Non-CG Methylation in the Human Genome. *Annual Review of Genomics and Human Genetics*, 16(1):55–77, 2015. doi: 10.1146/annurev-genom-090413-025437.
21. Francisco E. Baralle and Jimena Giudice. Alternative splicing as a regulator of development and tissue identity. *Nature Reviews Molecular Cell Biology*, 18(7):437–451, July 2017. ISSN 1471-0080. doi: 10.1038/nrm.2017.27.
22. Justin J. L. Wong, William Ritchie, Olivia A. Ebner, Matthias Selbach, Jason W. H. Wong, Yizhou Huang, Dadi Gao, Natalia Pinello, Maria Gonzalez, Kinsha Baidya, Annora Thoeng, Teh-Liane Khoo, Charles G. Bailey, Jeff Holst, and John E. J. Rasko. Orchestrated Intron Retention Regulates Normal Granulocyte Differentiation. *Cell*, 154(3):583–595, August 2013. ISSN 0092-8674. doi: 10.1016/j.cell.2013.06.052.
23. Yoseph Barash, John A. Calarco, Weijun Gao, Qun Pan, Xinchen Wang, Ofer Shai, Benjamin J. Blencowe, and Brendan J. Frey. Deciphering the splicing code. *Nature*, 465(7294): 53–59, May 2010. ISSN 1476-4687. doi: 10.1038/nature09000.
24. Hui Y. Xiong, Babak Alipanahi, Leo J. Lee, Hannes Bretschneider, Daniele Merico, Ryan K. C. Yuen, Yimin Hua, Serge Gueroussov, Hamed S. Najafabadi, Timothy R. Hughes, Quaid Morris, Yoseph Barash, Adrian R. Krainer, Nebojsa Jojic, Stephen W. Scherer, Benjamin J. Blencowe, and Brendan J. Frey. The human splicing code reveals new insights into the genetic determinants of disease. *Science*, 347(6218):1254806, January 2015. ISSN 0036-8075, 1095-9203. doi: 10.1126/science.1254806.
25. Yarden Katz, Eric T. Wang, Edoardo M. Airoldi, and Christopher B. Burge. Analysis and design of RNA sequencing experiments for identifying isoform regulation. *Nature Methods*, 7(12):1009–1015, December 2010. ISSN 1548-7105. doi: 10.1038/nmeth.1528.
26. Yuanhua Huang and Guido Sanguinetti. BRIE: transcriptome-wide splicing quantification in single cells. *Genome Biology*, 18:123, June 2017. ISSN 1474-760X. doi: 10.1186/s13059-017-1248-5.
27. Paul D. Ross, Jacky Guy, Jim Selfridge, Bushra Kamal, Noha Bahey, K. Elizabeth Tanner, Thomas H. Gillingwater, Ross A. Jones, Christopher M. Loughrey, Charlotte S. McCarroll, Mark E. S. Bailey, Adrian Bird, and Stuart Cobb. Exclusive expression of MeCP2 in the nervous system distinguishes between brain and peripheral Rett syndrome-like phenotypes. *Human Molecular Genetics*, 25(20):4389–4404, October 2016. ISSN 0964-6906. doi: 10.1093/hmg/ddw269. Publisher: Oxford Academic.
28. Daehwan Kim, Joseph M. Paggi, Chanhee Park, Christopher Bennett, and Steven L. Salzberg. Graph-based genome alignment and genotyping with HISAT2 and HISAT-genotype. *Nature Biotechnology*, 37(8):907–915, August 2019. ISSN 1546-1696. doi: 10.1038/s41587-019-0201-4.
29. Cole Trapnell, Brian A. Williams, Geo Pertea, Ali Mortazavi, Gordon Kwan, Marijke J. van Baren, Steven L. Salzberg, Barbara J. Wold, and Lior Pachter. Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. *Nature Biotechnology*, 28(5):511–515, May 2010. ISSN 1546-1696. doi: 10.1038/nbt.1621.
30. Yang Liao, Gordon K. Smyth, and Wei Shi. featureCounts: an efficient general purpose program for assigning sequence reads to genomic features. *Bioinformatics*, 30(7):923–930, April 2014. ISSN 1367-4803. doi: 10.1093/bioinformatics/btt656.
31. Michael I. Love, Wolfgang Huber, and Simon Anders. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biology*, 15(12):550, December 2014. ISSN 1474-760X. doi: 10.1186/s13059-014-0550-8.
32. W. J. Kent, A. S. Zweig, G. Barber, A. S. Hinrichs, and D. Karolchik. BigWig and BigBed: enabling browsing of large distributed datasets. *Bioinformatics*, 26(17):2204–2207, September 2010. ISSN 1367-4803. doi: 10.1093/bioinformatics/btq351.
33. David J. C. MacKay. Bayesian Interpolation. *Neural Computation*, 4(3):415–447, May 1992. ISSN 0899-7667. doi: 10.1162/neco.1992.4.3.415.
34. Michael E. Tipping. Sparse Bayesian Learning and the Relevance Vector Machine. *Journal of Machine Learning Research*, 1(Jun):211–244, 2001. ISSN ISSN 1533-7928.
35. Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, Jake Vanderplas, Alexandre Passos, David Cournapeau, Matthieu Brucher, Matthieu Perrot, and Édouard Duchesnay. Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research*, 12:2825-2830, October 2011. ISSN 1533-7928.
36. Leo Breiman. Random Forests. *Machine Learning*, 45(1):5–32, October 2001. ISSN 1573-0565. doi: 10.1023/A:1010933404324.

Figures

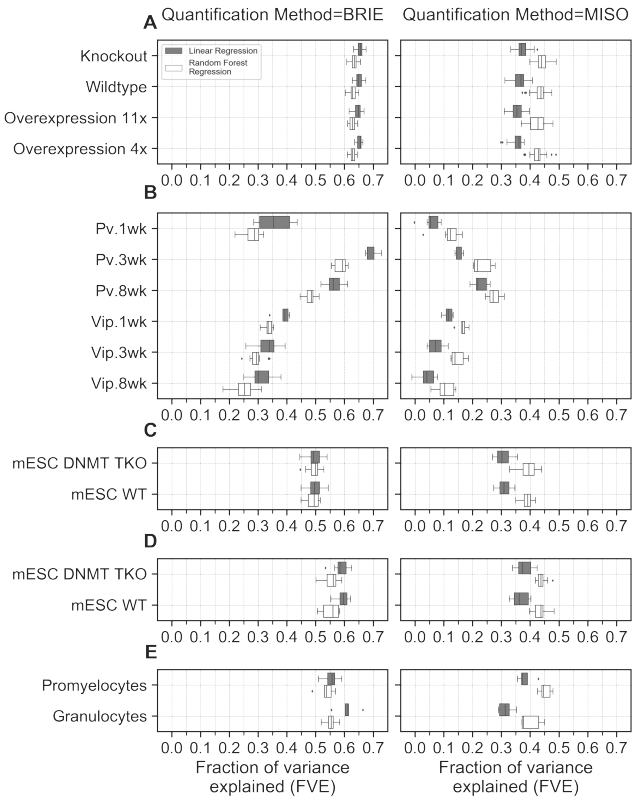

Fig. 1. Fraction of variance explained by regressing sequence features against splicing ratios. (A) Cultured human neurons expressing multiple levels of MeCP2(9) **(B)** *DNMT-TKO* and *WT* mESCs(18) **(C)** *DNMT-TKO* and *WT* mESCs(17) **(D)** *Pv* and *Vip* neurons at 1 week, 3 weeks and 8 weeks(19) **(E)** Promyelocytes and Granulocytes(14, 22). See Table 1 for detailed information about data-sets.

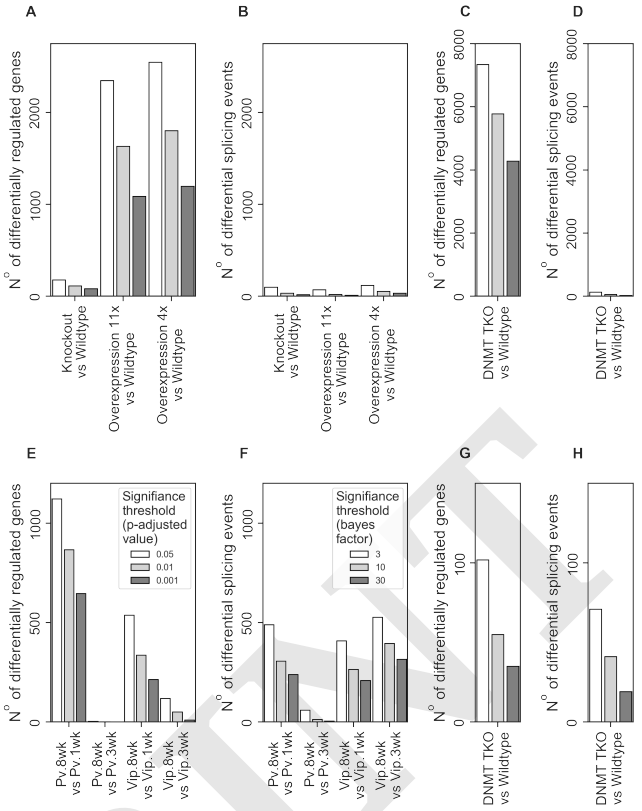

Fig. S1. Number of differential gene expression and splicing events across data sets. (A) Number of differentially regulated genes and **(B)** Number of differential splicing events in cultured neurons expressing multiple levels of MeCP2(9) **(C)** Number of differentially regulated genes and **(D)** Number of differential splicing events in *DNMT-TKO* mESCs(18) **(E)** Number of differentially regulated genes and **(F)** Number of differential splicing events in developing mouse neurons(19) **(G)** Number of differentially regulated genes and **(H)** Number of differential splicing events in *DNMT-TKO* mESCs(17)

PREPRINT

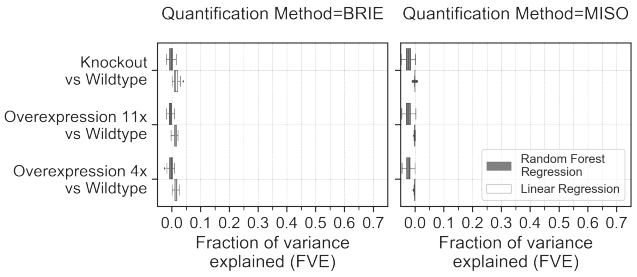

Fig. 2. Fraction of variance explained by regressing splicing variation against sequence features along with MeCP2 binding in cultured human neurons(9).

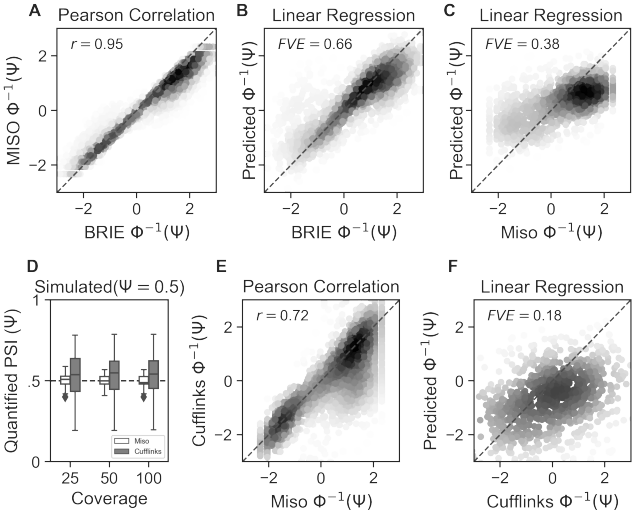

Fig. S2. Modelling transformed splicing ratios $y = \Phi^{-1}(\Psi)$, estimated using different quantification methods, against sequence features X. (A) Representative example of correlation between quantification methods BRIE and Miso. **(B)** Representative scatter plot showing prediction of splicing ratios using linear regression model learned from BRIE's quantification and sequence features. **(C)** Representation scatter plot showing prediction of splicing ratios using linear regression model learned from Miso's quantification and sequence features. **(D)** Comparison of splicing ratio estimation of known Ψ from simulated data at different coverage. **(E)** Representative example of correlation between quantification methods Cufflinks and Miso. **(F)** Representative scatter plot showing prediction of splicing ratios using linear regression model learned from Cufflinks' quantification and sequence features.

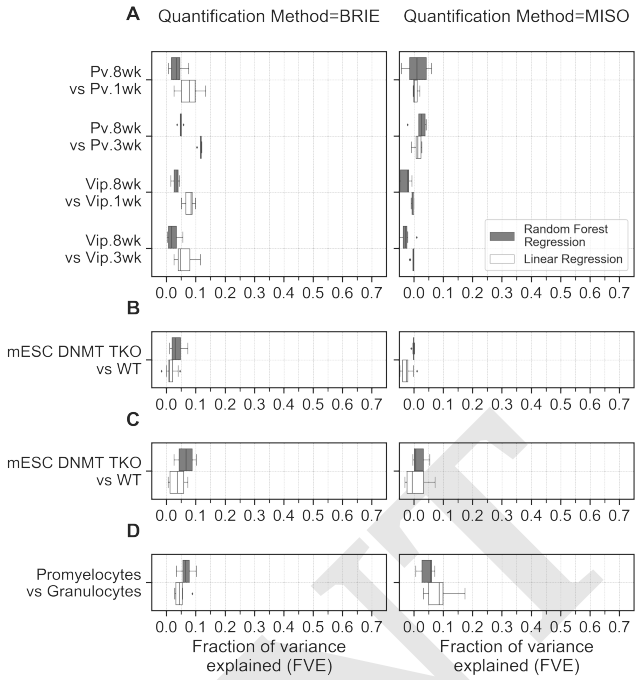

Fig. 3. Fraction of variance explained by regressing splicing variation against sequence features along with DNA methylation features. (A) *Pv* and *Vip* neurons at 1 week, 3 weeks and 8 weeks(19) **(B)** *DNMT-TKO* and *WT* mESCs(18) **(C)** *DNMT-TKO* and *WT* mESCs(17) **(D)** Promyelocytes and Granulocytes(14, 22).

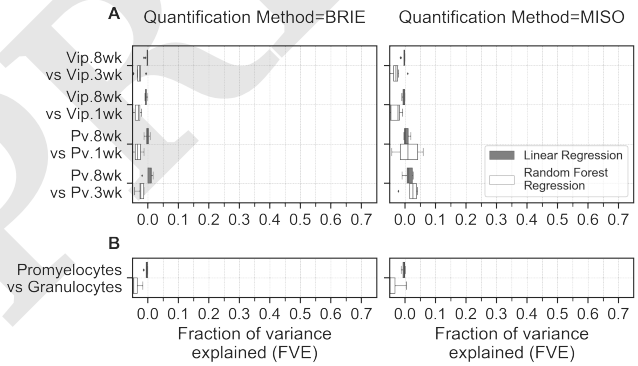

Fig. S3. Fraction of variance explained by regressing DNA methylation features against differential splicing ratios in (A) developing mouse neurons(19) **(B)** Promyelocytes and Granulocytes(14, 22)

PREPRINT

Materials and methods

Sequencing data sets. Table 1 details the data sets used in the analysis. These include RNA sequencing (RNA-seq) and Bilsulfite sequencing (BS-seq) libraries quantifying DNA methylation levels in *Dnmt1/3a/3b* triple-knockout (*DNMT-TKO*) and wildtype (*WT*) mouse embryonic stem cells (mESCs). Also, data sets concerning MeCP2 include RNA-seq and Chromatin immunoprecipitation followed by sequencing (ChIP-seq) libraries from post-mitotic (day 9 of differentiation) human neurons expressing multiple levels of MeCP2 and BS-seq libraries quantifying DNA methylation patterns from wildtype MeCP2-expressing neurons; RNA-Seq and BS-Seq libraries from mouse neuronal subtypes paravalbumin (*Pv*-) and vasoactive-intestinal-peptide (*Vip*-) expressing interneurons in mouse brain at different post-natal ages; RNA-seq and BS-seq libraries from cells of the mouse myeloid lineage, namely Granulocytes and Promyelocytes.

Table 1. Sequencing data sets considered for multi-variate quantitative analysis

Accession	Source	Type	Organism
GSE125660(9)	MeCP2 *WT* Neurons	BS-seq, RNA-seq, ChIP-seq	*Homo sapiens*
	MeCP2 *KO* Neurons	RNA-seq, ChIP-seq	
	MeCP2 *OE4x* Neurons		
	MeCP2 *OE11x* Neurons		
GSE64910(17)	*WT* mESCs	BS-seq, RNA-seq	*Mus musculus*
	DNMT-TKO mESCs	RNA-seq	
GSE67867(18)	*WT* mESCs		
	DNMT-TKO mESCs		
GSE103214(19)	*Pv* Neurons (1 week)	BS-seq, RNA-seq	
	Pv Neurons (3 weeks)		
	Pv Neurons (8 weeks)		
	Vip Neurons (1 week)		
	Vip Neurons (3 weeks)		
	Vip Neurons (8 weeks)		
GSE48307(22),GSE85517(14)	Promyelocytes	BS-seq, RNA-seq, ChIP-seq	
	Granulocytes		

Quantification of Splicing Ratios. RNA-Seq data sets are downloaded and sequencing reads are mapped to the reference genome using HISAT2(28) aligner. Splicing ratios are quantified using Miso(25), BRIE(26) and Cufflinks(29). We found Cufflinks to perform poorly in simulated data with known Ψ values (Figure S2D) when the sequence reads were single-end. The Pearson correlation between Cufflinks and Miso quantification in our data sets reflects this (representative scatter plot Figure S2E). Because two of the data sets that we have considered are sequenced with single-end reads and the resultant suboptimal modelling (Figure S2F), we decided to omit Cufflinks quantifications and regression models from further analyses.

Differential Gene Expression. Gene counts are extracted using featureCounts(30) from alignment files. Subsequently, differential gene expression is estimated using DESeq2(31).

Differential Splicing. Differential splicing is estimated using Miso(25).

Quantification of DNA methylation patterns. Processed BS-Seq data sets described in Table 1 are downloaded and DNA methylation ratio (mC basecalls to the count of all reads C as $\hat{m}_C = \frac{mC}{C}$) at individual nucleotide for CG and CA dinucleotides is accumulated as BigWig(32) tracks. DNA methylation is calculated in two metrics: 1) Region-specific mean methylation defined as $\hat{m}_C = \frac{\sum m_C}{N_C}$ where N_C is the number of C's within the region 2) Region-specific methylation density defined as $\hat{M}_C = \frac{N_C \cdot \hat{m}_C}{L}$ where L is the length of the region. These two metrics are calculated for all the genes such that every gene has separate quantifications for CA and CG dinucleotides. For each gene, several specific regions are considered 1) ASE (alternative spliced exon) 2) combination of all exons containing the ASE 3) combination of all exons without the ASE 4) combination of all introns 5) combination of all introns without the ASE.

Quantification of ChIP-seq signal. Processed ChIP-seq data sets described in Table 1 are downloaded and $log_2 \frac{ChIP}{Input}$ signal is calculated using `bigWigAverageOverBed`(32) for regions around the alternative exon.

Regressing Splicing Ratios with Sequence Features. First, we use an inverse `probit` transformation of our splicing ratios Ψ i.e, $y = \Phi^{-1}(\Psi)$ to map splicing ratios from the interval [0,1] to the whole of the real numbers. Subsequently, we model the transformed splicing ratio y as a function of a set of j covariates (sequence features) $X \in \mathbb{R}^j$. We split the data, fit

PREPRINT

the model and compute the coefficient of determination (FVE) which is analogous to R^2, 5 consecutive times (with different splits each time) to avoid overfitting. The FVE calculated is the amount of variation in the test set that can be explained by the covariates (sequence features). $FVE(y,\hat{y}) = 1 - \frac{\sum_{i=0}^{n}(y_i-\hat{y}_i)^2}{\sum_{i=0}^{n}(y_i-\bar{y})^2}$ where $\bar{y} = \frac{1}{n}\sum_{i=0}^{n} y_i$. This is applied for both regression approaches.

Linear Regression. We model the y (transformed splicing ratio) as a linear function of X (sequence features). To regularize the weight and noise parameters in the regression problem, we follow the regularization approach(33, 34) which is implemented in `scikit-learn`(35) package.

Random Forest Regression. We generate random decision trees that are fitted on sets of training data from X (sequence features) and y (transformed splicing ratio) to form a meta estimator known as random forests(36), implemented in `scikit-learn`(35) package. Using averaging from a number of trees, we predict the transformed splicing ratio y on the test data.

Regressing Differential Splicing Ratios with Differential DNA methylation and Sequence Features. As discussed in the previous section, we use the same approach for the modelling. In this instance, we add more covariates in the form of DNA methylation features. For mean methylation metric, we use an inverse `probit` transformation and take the difference.

$$\Delta\Phi^{-1}(\hat{m}_C) = \Phi^{-1}(\hat{m}_{C_t}) - \Phi^{-1}(\hat{m}_{C_c})$$

where $t = treatment/time-point$ and $c = control/time-point$. For methylation density metric, we calculate the difference directly.

$$\Delta\hat{M}_C = \hat{M}_{C_t} - \hat{M}_{C_c}$$

We calculate the differential splicing ratio i.e, $\Delta y = \Delta\Phi^{-1}(\Psi)$

$$\Delta\Phi^{-1}(\Psi) = \Phi^{-1}(\Psi_t) - \Phi^{-1}(\Psi_c)$$

and model it to a set of covariates which now consists of sequence features and DNA methylation feautures.

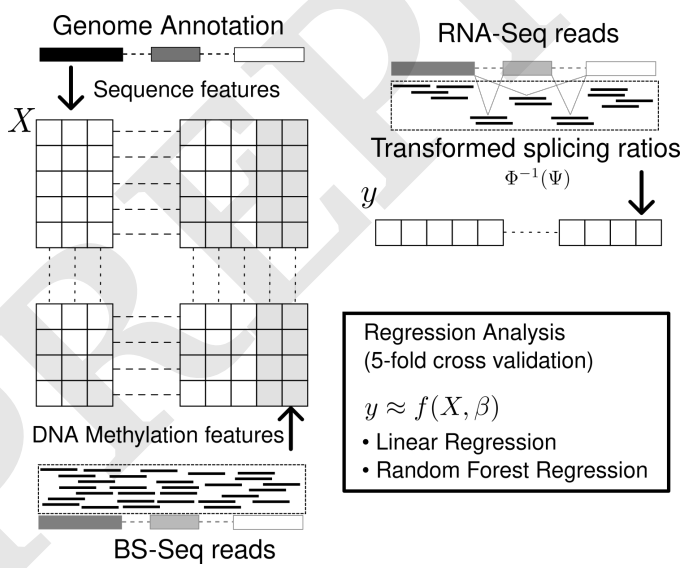

Fig. S4. Graphical representation of the overall regression analyses. Accumulation of sequence features from genome annotation, accumulation of DNA Methylation features from BS-Seq (Bisulfite Sequencing) reads and quantification of Splicing Ratios (Ψ) from RNA-Seq (RNA Sequencing) reads.

Regression Analysis Overview.

PREPRINT